Speech emotion recognition with TGI+.2 classifier

Julia Sidorova
Universitat Pompeu Fabra
Barcelona, Spain
julia.sidorova@upf.edu

Abstract

We have adapted a classification approach coming from optical character recognition research to the task of speech emotion recognition. The classification approach enjoys the representational power of a syntactic method and efficiency of statistical classification. The syntactic part implements a tree grammar inference algorithm. We have extended this part of the algorithm with various edit costs to penalise more important features with higher edit costs for being outside the interval, which tree automata learned at the inference stage. The statistical part implements an entropy based decision tree (C4.5). We did the testing on the Berlin database of emotional speech. Our classifier outperforms the state of the art classifier (Multilayer Perceptron) by 4.68% and a baseline (C4.5) by 26.58%, which proves validity of the approach.

1 Introduction

In a number of applications such as human-computer interfaces, smart call centres, etc. it is important to be able to recognise people's emotional state. An aim of a speech emotion recognition (SER) engine is to produce an estimate of the emotional state of the speaker given a speech fragment as an input. The standard way to do SER is through a supervised machine learning procedure (Sidorova et al., 2008). It also should be noted that a number of alternative classification strategies has been offered recently, such as unsupervised learning (Liu et al., 2007) and numeric regression (Grimm et al., 2007) etc, and which are preferable under certain conditions.

Our contribution is a new algorithm of a mixed design with syntactic and statistical learning, which we borrowed from optical character recognition (Sempere, Lopez, 2003), extended, and adapted for SER. The syntactic part implements tree grammar inference (Sakakibara, 1997), and the statistical part implements C4.5 (Quinlan, 1993). The intuitive reasons underlying this solution are as follows. We would like to have a classification approach that enjoys the representational power of a syntactic method and efficiency of statistical classification. First we model the objects by means of a syntactic method, i.e. we map samples into their representations. A representation of a sample is a set of seven numeric values, signifying to which degree a given sample resembles the averaged pattern of each of seven classes. Second, we learn to classify the mappings of samples, rather than feature vectors of samples, with a powerful statistical method. We called the classifier *TGI+*, which stands for Tree Grammar Inference and the plus is for the statistical learning enhancement. In this paper we present the second version of TGI+, which extends TGI+.1 (Sidorova et al., 2008) and the difference is that we have added various edit costs to penalise more important features with higher edit costs for being outside the interval, which tree automata learned at the inference stage. We evaluated TGI+ against a state of the art classifier. To obtain a state of the art performance, we constructed a speech emotion recogniser, following the classical supervised learning approach with a top performer out of more than 20 classifiers from the weka package, which turned out to be multilayer perceptron (MLP) (Witten, Frank, 2005). Experimental results showed that TGI+ outperforms MLP by 4.68%.

The structure of this paper is as follows: in this section below we explain construction of a classical speech emotion recognizer, in Section 2 we explain TGI+; Section 3 reports testing results for both, the state of the art recogniser and TGI+. Section 4 and 5 is discussion and conclusions.

Proceedings of the EACL 2009 Student Research Workshop, pages 54–60,
Athens, Greece, 2 April 2009.

1.1 Classical Speech Emotion Recogniser

A classical speech emotion recognizer is comprised of three modules: Feature Extraction, Feature Selection, and Classification. Their performance will serve as a baseline for TGI+ recognizer.

1.1.1 Feature Extraction and Selection

In the literature there is a consensus that global statistics features lead to higher accuracies compared to the dynamic classification of multivariate time-series (Schuller et al., 2003). The feature extraction module extracts 116 global statistical features, both prosodic and segmental, a full list and explanations for which can be found in (Sidorova, 2007).

The feature selection module implements a wrapper approach with forward selection (Witten, Frank, 2005) in order to automatically select the most relevant features extracted by the previous module.

1.1.2 Classification

The classification module takes an input as a feature vector created by the feature selector, and applies the Multilayer Perceptron classifier (MLP) (Witten, Frank, 2005), in order to assign a class label to it. The labels are the emotional states to discriminate among. For our data, MLP turned out to be the top performer among more than 20 other different classifiers; details of this comparative testing can be found in (Sidorova, 2007).

2 TGI+ classifier

The organisation of this section is as follows. In paragraph 2.1 we explain the TGI+.1 classifier and show how its parts work together. TGI+.2 is an extension of TGI+.1 and we explain it right afterwards. In paragraph 2.2 we briefly remind the C4.5 algorithm. Further in the paper in paragraph 4.1 we show that our TGI+ algorithm was correctly constructed and that we arrived to a meaningful combination of methods from different pattern recognition paradigms.

2.1 TGI+

TGI+.1 is comprised of four major steps we explain below. Fig 1 graphically depicts the procedure.

Step 1: In order to perform tree grammar inference we represent samples by tree structures. Divide the training set into two subsets T_1 (39% of training data) and T_2 (the rest of training data). Utterances from T_1 are converted into tree structures, the skeleton of which is defined by the grammar below. S denotes a start symbol of the formal grammar (in the sense of a term-rewriting system):
$\{S \longrightarrow$ ProsodicFeatures SegmentalFeatures;
ProsodicFeatures $\longrightarrow$ Pitch Intensity Jitter Shimer;
SegmentalFeatures $\longrightarrow$ Energy Formants;
Pitch $\longrightarrow$ Min Max Quantile Mean Std MeanAbsoluteSlope;
etc. $\}$

The *etc.* stands for further terminating productions, i.e. the productions which have low level features on their right hand side. All trees have 116 leaves, each corresponding to one of the 116 features from the sample feature vector. We put trees from one class into one set. In our dataset we have the following seven classes to recognise among: fear, disgust, happiness, boredom, neutral, sadness and anger. Therefore, we have seven sets of trees. We put trees from one class into one set.

Step 2: Apply tree grammar inference to learn seven automata accepting a different type of emotional utterance each. Grammar inference is a method to learn a grammar from examples. In our case, it is *tree* grammar inference, because we deal with trees representing utterances. The result of this step is seven automata, one for each of seven emotions to be recognised.

Step 3: Calculate edit distances between obtained tree automata and trees in the training set. Edit distances are then calculated between each automaton obtained at step two and each tree representing utterances from the training set ($T_1 \cup T_2$). The calculated edit distances are put into a matrix of size: (cardinality of the training set) $\times$ 7 (the number of classes).

Step 4: Run C4.5 over the matrix to obtain a decision tree. The C4.5 algorithm is run over this matrix in order to obtain a decision tree, classifying each utterance into one of the seven emotions, according to edit distances between a given utterance and the seven tree automata. The accuracies obtained from testing this decision tree are the accuracies of TGI+.1.

TGI+.2 Our extension of the algorithm as proposed in (Sempere, Lopez, 2003) has to do with Step 3. In TGI+.1 all edit costs equated to 1. In

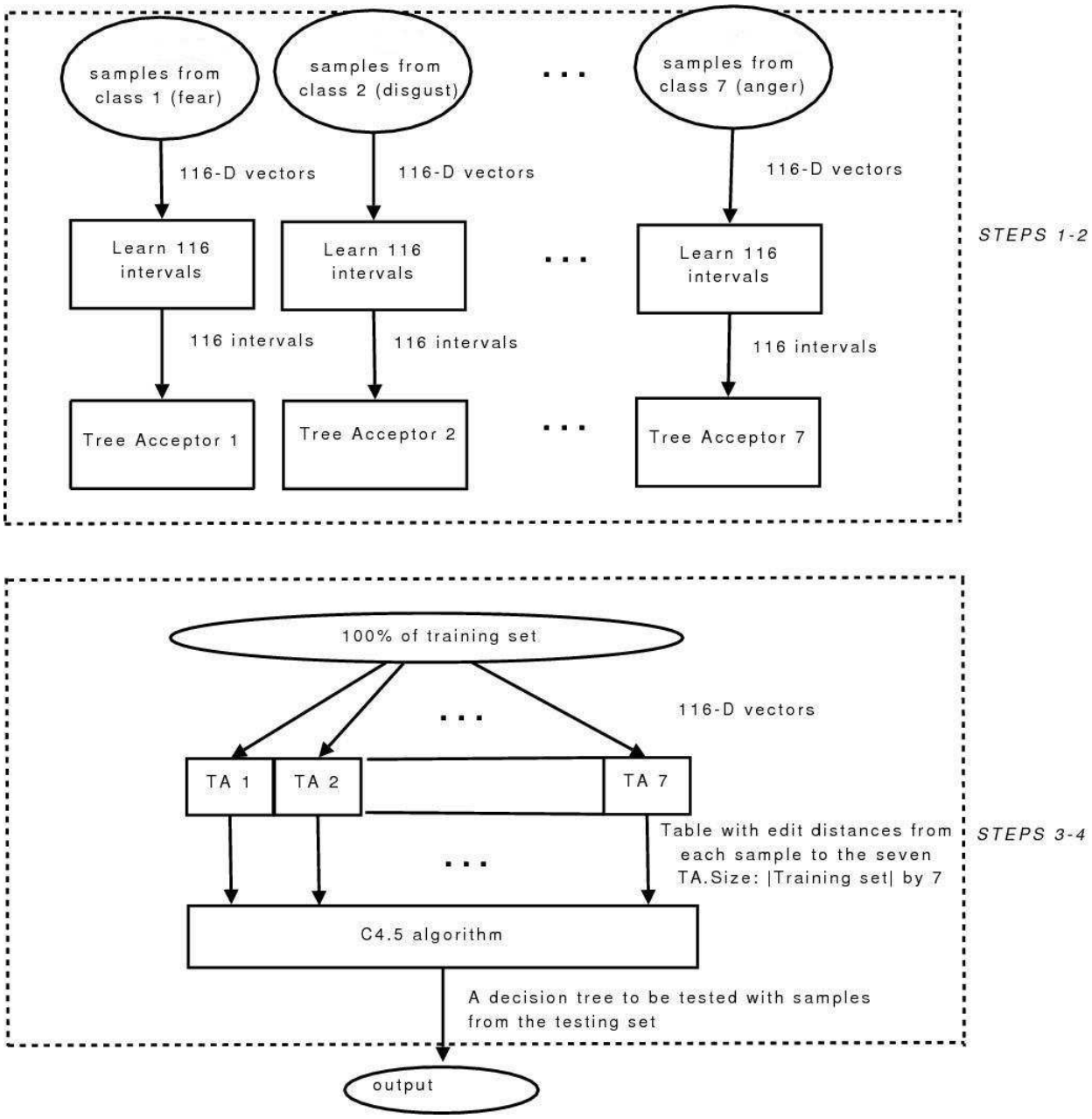

Figure 1: **TGI+ steps.** Step 1: In order to perform tree grammar inference, represent samples by tree structures. Step 2: Apply tree grammar inference to learn seven automata accepting a different type of emotional utterance each. Step 3: Calculate edit distances between obtained tree automata and trees in the training set. While calculating edit distances, penalise more important features with higher costs for being outside its interval. The set of such features is determined exclusively for every class through a feature selection procedure. Step 4: Run C4.5 over the matrix to obtain a decision tree.

other words, if a feature value fits the interval a tree automaton has learned for it, the acceptance cost of the sample is not altered. If a feature value is outside the interval the automaton has learnt for it, the acceptance cost of the sample processed is incremented by one. In TGI+.2 some edit costs have a coefficient greater than 1 (1.5 in the current version). In other words, more important features are penalised with higher costs for being outside its interval. The set of these more important features is determined exclusively for every class (anger, neutral, etc.) through a feature selection procedure. The feature selection procedure implements a wrapper approach with forward selection.

Concluding the algorithm description, let us explain how TGI+ classifiers an input sample, which is fed to the automata in the form a 116 dimensional feature vector. Firstly TGI+ calculates distances from the sample to seven tree automata (the automata learnt 116 feature intervals at the inference step). Secondly TGI+ uses the C 4.5 decision tree to classify the sample (the decision tree was learnt from the table, where distances to seven automata to all the training samples had been put).

2.2 C4.5 Learning algorithm

C4.5 belongs to the family of algorithms that employ a topdown greedy search through the space of possible decision trees. A decision tree is a representation of a finite set of if-then-else rules. The main characteristics of decision trees are the following:

1. The examples can be defined as a set of numerical and symbolic attributes.
2. The examples can be incomplete or contain noisy data.
3. The main learning algorithms work under Minimum Description Length approaches.

The main learning algorithms for decision trees were proposed by Quinlan (Quinlan, 1993). First, he defined ID3 algorithm based on the information gain principle. This criterion is performed by calculating the entropy that produces every attribute of the examples and by selecting the attributes that save more decisions in information terms. C4.5 algorithm is an evolution of ID3 algorithm. The main characteristics of C4.5 are the following:

1. The algorithm can work with continuous attributes.
2. Information gain is not the only learning criterion.
3. The trees can be post-pruned in order to refine the desired output.

3 Experimental work

We did the testing on acted emotional speech from the Berlin database (Burkhardt el al., 2005). Although acted material has a number of well known drawbacks, it was used to establish a proof of concept for the methodology proposed and is a benchmark database for SER. In the future work we plan to do the testing on real emotions. The Berlin Emotional Database (EMO-DB) contains the set of emotions from the MPEG-4 standard (anger, joy, disgust, fear, sadness, surprise and neutral). Ten German sentences of emotionally undefined content have been acted in these emotions by ten professional actors, five of them female. Throughout perception tests by twenty human listeners 488 phrases have been chosen that were classified as more than 60% natural and more than 80% clearly assignable. The database is recorded in 16 bit, 16 kHz under studio noise conditions.

As for the testing protocol, 10-fold cross-validation was used. Recall, precision and F measure per class are given in Tables 3, 4.1 and 4.2 for C4.5, MLP and TGI+, respectively. The overall accuracy of MLP, the state of the art recogniser, is 73.9% and the overall accuracy of the TGI+ based recogniser is 78.58%, which is a 4.68% $\pm$ 3.45% in favour of TGI+. The confidence interval was calculated as follows: $Z\sqrt{\frac{p(1-p)}{n}}$, where p is accuracy, n is cardinality of the data set, and Z is a constant for the confidence level of 95%, i.e. $Z = 1.96$. The proposed TGI+ has also been evaluated against C4.5 to find out which is the contribution of moving from the feature vector representation of samples to the distance to automata one. C4.5 performs with 52.9% of acuracy, which is 25.68% less than TGI+. The positive outcome of such contrastive testing in favour of TGI+ was expected, because TGI+ was designed to enjoy strengths of two paradigms: syntactic and statistical, while MLP (or C4.5) is a powerful single paradigm statistical method.

class	precision	recall	F measure
fear	0.49	0.44	0.46
disgust	0.26	0.24	0.26
happiness	0.35	0.36	0.35
boredom	0.49	0.55	0.52
neutral	0.51	0.46	0.49
sadness	0.71	0.82	0.76
anger	0.69	0.7	0.7

Table 1: Baseline recognition with C4.5 on the Berlin emotional database. The overall accuracy is 52.9%, which is 25.68% less accurate than TGI+.

class	precision	recall	F measure
fear	0.82	0.74	0.77
disgust	0.72	0.74	0.73
happiness	0.52	0.49	0.51
boredom	0.73	0.75	0.74
neutral	0.71	0.78	0.75
sadness	0.88	0.94	0.91
anger	0.75	0.76	0.75

Table 2: State of the art recognition with MLP on the Berlin emotional database. The overall accuracy is 73.9%, which is 4.68% less accurate than TGI+.

4 Discussion

4.1 Correctness of algorithm construction

While constructing TGI+, it is of critical importance that the following condition holds: *The accuracy of TGI+ is better than that of tree acceptors and C4.5.* If this condition holds, then TGI+ is well constructed. We tested TGI+, tree automata as acceptors and C4.5 on the same Berlin database under the same experimental settings. The tree automata and C4.5 perform with 43% and 52.9% of accuracy respectively, which is 35.58% and 25.68% worse than the accuracy of TGI+. Therefore the condition is met and we can state that we arrived to a meaningful combination of methods from different pattern recognition paradigms.

4.2 A combination of statistical and syntactic recognition

Syntactic recognition is a form of pattern recognition, where items are presented as pattern structures, which take account of more complex interrelationships between features than simple numeric feature vectors used in statistical classification. One way to represent such structure is strings

class	precision	recall	F measure
fear	0.66	0.66	0.66
disgust	0.6	0.6	0.6
happiness	0.86	0.73	0.81
boredom	0.81	0.72	0.77
neutral	0.64	0.79	0.71
sadness	0.83	0.83	0.83
anger	0.89	0.93	0.91

Table 3: Performance of the TGI+ based emotion recognizer on the Berlin emotional database. The overall accuracy is 78.58%.

(or trees) of a formal language. In this case differences in the structures of the classes are encoded as different grammars. In our case, we have numeric data in place of a finite alphabet, which is more traditional for syntactic learning. The syntactic method does the mapping of objects into their models, which can be classified more accurately than objects themselves.

4.3 Why tree structures?

Looking at the algorithm, it might seem redundant to have tree acceptors, when the same would be possible to handle with a finite state automaton (that accepts the class of regular string languages). Yet tree structures will serve well to add different weights to tree branches. The motivation behind is that acoustically some emotions are transmitted with segmental features and others with prosodic, e.g. prosody can be prioritised over segmental features or vice versa (see also Section 4.5).

4.4 Selection of C4.5 as a base classifier in TGI+

A natural question is: given that MLP outperforms C4.5, which are the reasons for having C4.5 as a base classifier in TGI+ and not the top statistical classifier? We followed the idea of (Sempere, Lopez, 2003), where C4.5 was the base classifier. We also considered the possibility of having MLP in place of C4.5. The accuracies dramatically went down and we abandoned this alternative.

4.5 Future work

I. Tuning parameters. There are two tuning parameters. To exclude the possibility of overfitting, the testing settings should be changed to the protocol with disjoint training, validation and testing sets. We have not done the experiments under the

new training/testing settings, yet we can use the old 10-f cross validation mode to see the trends. Tuning parameter 1 is the point of division of the training set into the two subsets T_1 and T_2, i.e. a division of the training data to train the statistical and syntactic classifier. The division point should be shifted from 5% for syntactic and 100% for statistical to 100% to train both syntactic and statistical models. The point of division of the training data is an accuracy sensitive parameter. Our rough analysis showed that the resulting function (point of division for abscissa and accuracy for ordinate) has a hill shape with one absolute maximum, and we made a division roughly at this point: 39% of the training data for the syntactic model. Finding the best division in fair experimental settings remains for future work.

Tuning parameter 2 is a set of edit costs assigned to different branches of the tree acceptors. A linguistic approach is an alternative to the feature selection we followed so far. This is the point at which finite state automata cease to be an alternative modelling device. The motivation behind is that acoustically some emotions are transmitted with segmental features and others with prosodic (Barra, et al., 1993). A coefficient of 1.5 on the prosodic branches brought 2% of improvement of recognition for boredom, neutral and sadness.

II. Testing TGI+ on authentic emotions. It has been shown that authentic corpora have very different distributions compared to acted speech emotions (Vogt, Andre, 2005). We must check whether TGI+ is also a top performer, when confronted with authentic corpora.

III. Complexity and computational time. A number of classifiers, like MLP (but not C4.5) require a prior feature selection step, while TGI+ always uses a complete set of features, therefore better accuracies come at the cost of higher computational complexity. We must analyse such advantages and disadvantages of TGI+ compared to other popular classifiers.

5 Conclusions

We have adapted a classification approach coming from optical character recognition research to the task of speech emotion recognition. The general idea was that we would like a classification approach to enjoy the representational power of a syntactic method and the efficiency of statistical classification. The syntactic part implements a tree grammar inference algorithm. The statistical part implements an entropy based decision tree (C4.5). We did the testing on the Berlin database of emotional speech. Our classifier outperforms state of the art classifier (Multilayer Perceptron) by 4.68% and a baseline (C4.5) by 26.58%, which proves validity of the approach.

6 Acknowledgements

This research was supported by AGAUR, the Research Agency of Catalonia, under the BE-DRG 2007 mobility grant. We would like to thank Lab of Spoken Language Systems, Saarland University, where much of this work was completed.

References

Barra R., Montero J.M., Macias-Guarasa, DHaro, L.F., San-Segundo R., Cordoba R. 2005. *Prosodic and segmental rubrics in emotion identification*. Proc. ICASSP 2005, Philadelphia, PA, March 2005.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B. 2002. *A database of German Emotional Speech.* Proc. Interspeech 2005, ISCA, pp 1517-1520, Lisbon, Portugal, 2005.

Grimm M., Kroschel K., Narayanan S. 2007. *Support vector regression for automatic recognition of spontaneous emotions in speech*, Proc. of ICASSP, Honolulu, Hawaii, April 2007.

Liu, J., Chen, C., Bu, J., You, M, Tao, J. 2007. *Speech emotion recognition using an enhanced co-training algorithm*, in Proc. of ICME, Bejing, China, July, 2007.

Lopez D., Espana, S. 2002. *Error-correcting tree-language inference*. Pattern Recognition Letters 23, pp. 1-12. 2002

Sakakibara, Y. 1997. *Recent advances of grammatical inference.* Theoretical Computer Science 185, pp. 15-45. Elsevier. 1997.

Schuller B., Rigoll G. Lang M. 2003. *Hidden Markov Model-Based Speech Emotion Recognition*, Proc. of ICASSP 2003, Vol. II, pp. 1-4, Hong Kong, China, 2003.

Sempere J. M., Lopez D. 2003. *Learning decision trees and tree automata for a syntactic pattern recognition task.* Pattern Recognition and Image Analysis. Lecture notes in CS. Berlin. Volume 2652. pp. 943-950, 2003.

Sidorova J. 2007. *DEA report: Speech Emotion Recognition*. Appendix 1 (for the feature list) and Section 3.3. (for a comparative testing of various weka classifiers) . http://www.glicom.upf.edu/tesis/sidorova.pdf Universitat Pompeu Fabra

Sidorova J., McDonough J., Badia T. 2008. *Automatic Recognition of Emotive Voice and Speech*, in (Eds.) K. Izdebski. Emotions in The Human Voice, Vol. 3, Chap. 12, Plural Publishing, San Diego, CA, 2008.

Quinlan, J.R. 1993. *C4.5: Programs For Machine Learning*. Morgan Kaufmann, Los Altos. 1993.

Vogt, T. Andre, E. 2005. *Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition.* Proc. ICME 2005, Amsterdam, Netherlands, 2005.

Witten I.H., Frank E. 2005. Sec. 7.1 (for feature selection) and Sec. 10.4 (for multilayer perceptron) in *Data Mining. Practical Machine Learning Tools and Techniques.* Elsevier. 2005.