The recognition of emotions from speech using GentleBoost classifier. A comparison approach

Dragoş Datcu, Leon J.M. Rothkrantz

Abstract: *The recognition of the internal emotional state of one person plays an important role in several human-related fields. Among them, human-computer interaction has recently received special attention. The current research is aimed at the analysis of segmentation methods and of the performance of the GentleBoost classifier on emotion recognition from speech. The data set used for emotion analysis is Berlin - a database of German emotional speech. A second data set is DES – Danish Emotional Speech data set is used for comparison purposes. Our contribution for the research community consists in a novel extensive study on the efficiency of using distinct numbers of frames per speech utterance for emotion recognition. Eventually, a set of GentleBoost 'committees' with optimal classification rates is determined based on an exhaustive study on the generated classifiers and on different types of segmentation.*

Key words: *Emotion recognition, speech processing, GentleBoost, Human-Computer Interfaces.*

INTRODUCTION

As the recent developments on speech driven technologies has led to even more reliable human computer systems, there has been an increasing interest in studying more sophisticated techniques to handle the emotional state of the speaker. The quality of the interaction between human beings and computers greatly improves by providing methods to automatically perceive and generate the feedback based on human non-verbal communication. In this context, the attempt to model the user's emotions by examining the acoustic cues has become a relevant research topic.

The paper shows a research on using different multi frame speech segmentation techniques on emotion recognition. The classifier chosen to model the emotion characteristics in speech is based on Gentle AdaBoost method for a maximum 200 training steps. The optimal classifiers are determined by employing ROC graphs to show the trade-off between the hit and the false positive rates. One important issue for the recognition of emotions in speech represents the segmentation of the speech signal. The way this process is done dramatically affects the subsequent results of the recognition of emotions. The current paper presents an attempt to study different approaches on speech signal segmentation. Given the set of prosodic features, we determine the segmentation type and the utterance frame structure that leads to good recognition of emotions. Finally, the overall recognition results are analysed against each type of segmentation for two different data sets.

PREVIOUS WORK

Recently, the recognition of emotions in speech has been extensively researched and various methods have been used. For example, Yu and al. [10] applied a multilevel structure based on coupled hidden Markov models to estimate engagement levels in continuous natural speech. The continuous speech signal is segmented into spoken utterances and the acoustic features are computed from each utterance portion. The extracted non-linguistic information is used for predicting the emotional states such as discrete emotion types or arousal/valence levels by employing SVM-based classifiers. The HMM uses the previous information to model the user's emotional state and engagement in conversation as a dynamic, continuous process. Chateau at al. [4] presents a study of the perception, the analysis and the modelling of styles or the 'emotional quality' of speech. The speech emotional quality is evaluated in terms of the emotional content that describes the listener's global impressions as elicited by their audition. Specific subjective

The research reported here is part of the Interactive Collaborative Information Systems (ICIS) project, supported by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024.

criteria for evaluating the emotional quality are used to generate perceptive portraits of the speech. The evaluation is carried by using linear models to connect the perceptive portraits to physical data derived from signal analysis. Some work has also been focussed on using additional information regarding speech. The paper of [8] uses three sources of information - acoustic, lexical and discourse - for recognizing emotions. Linear discriminant and k-nearest neighbourhood classifiers are used to classify acoustical information to anger and frustration - as negative emotions and to neutral or positive emotions. The different features are extracted by using certain portions of the signal. A noticeable approach stands for multimodal analysis that aims at improving the recognition rates for the emotional state by fusing the results on separate modalities. The advantage of such methods relates to the overcoming the limited information that can be gathered from each single modality. The work of [3] analysis the strengths and the limitations of systems based on the fusion of facial expression and acoustical information analysis at the decision level and in the case of feature level integration. Kwon et al. [7] provides a comparison on the emotion recognition performance of various classifiers. They obtained SVM and HMM based classifiers with significantly better results on SUSAS database from the previous approaches. A recent research of Rothkrantz et al. [9] focuses on studying the effect of the workload on speech production by making use of a psychological experimental setup. A full analysis on each acoustic feature is conducted in order to create efficient models for stress detection.

MODEL

Before being actually used for analysis, the speech signal has to undergo a set of operations. The first operation is filtering for noise reduction. The gender of the speakers is taken into account for creating separate data sets for the training and testing operations. Further on, the segmentation operation is run and different data sets are obtained, depending on the number of frames per utterance and the frame configuration. The frame configuration indicates which frames are selected in each utterance for extracting the acoustic features. The result consists in a set of prosodic feature values that represent the original data. These values are further used in parametric classification of emotion in speech.

Data sets

The first data set used for emotion analysis from speech is Berlin [2] – a database of German emotional speech. The database contains utterances of both male and female speakers, two sentences. The emotions were simulated by ten native German actors (five female and five male). The result consists of ten utterances (five short and five long sentences). The length of the utterance samples ranges from 1.2255 seconds to 8.9782 seconds. The recording frequency is 16kHz. The final speech data set contains the utterances for which the associated emotional class was recognized by at least 80\% of the listeners. Following a speech sample selection, an initial data set was generated comprising 456 samples and six basic emotions (*anger*: 127 samples, *boredom*: 81 samples, *disgust*: 46 samples, *anxiety/fear*: 69 samples, *happiness*: 71 samples and *sadness*: 62 samples). Subsequently, the DES database [5] is used for comparison. The DES database contains five emotions under investigation (neutral, surprise, happiness, sadness, anger). In order to record the emotion enriched speech signals two male and two female actors were used. Each of the four actors had to speak several utterances once for each of the five emotions. The utterances involved 2 single words, 9 sentences and 2 passages of fluent speech. In addition, there are 8 passages and 10 sentences for target voices. For mixed male and female utterances, the final data set for analysis has 279

samples with the following structure: (*anger*: 46 samples, *happy*: 48 samples, *neutral*: 46 samples, *sadness*: 46 samples, *surprise*: 48 samples, *targeted*: 45 samples).

Multi-frame analysis

In the case of emotion recognition from speech, the analysis is handled separately for different number of frames per utterance. In the current approach there are five types of splitting methods performed on initial data. Each type of splitting produces a number of data sets, according to all the frame combinations in one utterance. Although the analysis is done separately on *male*, *female* and *male and female* speakers, the current paper focuses only on mixed voices. For each of the three gender cases there is a number of 1065 data sets to be considered (table 1).

Table 1: The utterance segmentation and the number of resulting data sets.

Nr.of frames per utterance	1	2	3	5	10	Total
Nr. of data sets	1	3	7	31	1023	1065

Feature extraction

The Praat [1] tool was used for extracting the features from each sample from all generated data sets. According to each data set frame configuration, the parameters **mean**, **standard deviation**, **minimum** and **maximum** of the following acoustic features were computed: Fundamental frequency (pitch), Intensity, F1, F2, F3, F4 and Bandwidth. All these parameters form the input for separate GentleBoost classifiers according to data sets with distinct segmentation characteristics.

RESULTS

The GentleBoost *committee* is trained for a maximum number of 200 stages. Separate data sets containing male, female and both male and female utterances are considered for training and testing the classifier models. The performance of each classifier is evaluated with the 5-fold cross validation (for Berlin data set) and with 2-fold cross validation (for DES data set) methods. Depending on the number of sub-frames per speech frame, the different data sets are used to generate sets of classifiers. One curve on the graph stands for the set of representative GentleBoost strong classifiers generated by using the specific data set, associated with a certain split configuration. Each node on one curve relates to one classifier in the set. The ROC graph in figure 1 shows the trade-off between the hit and the false-positive rates for all the GentleBoost classifiers generated from Berlin data set. The correspondent ROC graph for DES data set is shown is figure 2. Each point on the figure stands for one GentleBoost classifier that is selected using the highest true-positive rate criterion. For each emotion class, a total number of 200 points is taken into account and only the ones with the highest scores are displayed on the same emotion curve. By analyzing each emotion curve separately, the final strong committee to be chosen is the one that is the closest to the north-west corner of the figure. In other words, the classifier in question is the one that has the highest true positive rate (*tpr*) while the false positive rate (*fpr*) is the lowest in the set of classifiers on the same curve.
Table 2 (for Berlin data set) and table 3 (for DES data set) depict the characteristics of each strong classifier that is selected for each emotion curve separately. The column *nr.stages* shows the number of stages required to train the associated strong committee. An additional field (*ac*) in each table shows the accuracy rate achieved by the classifiers. Each classifier is identified by the structure of the frames into the utterance sample

(column *frames*). A digit from one binary sequence specifies that the correspondent frame contributes ('1') or not ('0') with features at the classification process.

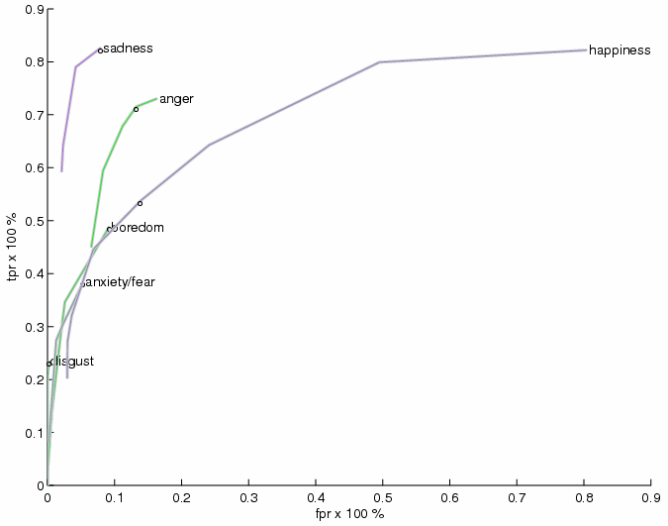

Figure 1: ROC graph that show the committees with the highest true positive rates for each emotion class. The classifiers are generated following the analysis on Berlin data set.

An observation on the tables proves that the majority of the strong classifiers lying on the emotion curves in the ROC graph clearly express the efficiency of using a ten frames per utterance configuration for the segmentation.

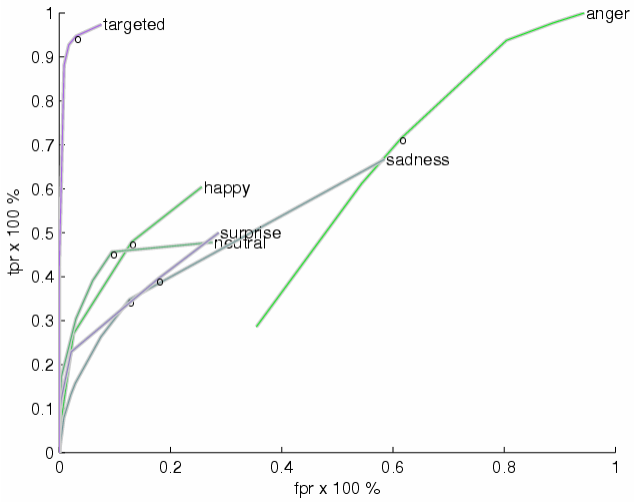

Figure 2: ROC graph that show the committees with the highest true positive rates for each emotion class. The classifiers are generated following the analysis on DES data set.

Due to the differences on the emotion classes for Berlin and DES data sets, it is rather hard to make comparisons on the performances achieved in the analysis. However, there are three common emotion classes: *anger*, *happiness* and *sadness*. The overall results indicate the higher performance of classifiers trained on Berlin data set over the classifiers trained on DES data set. This can be mainly explained by the bigger size of the training set in the case of Berlin data set. Although the true positive rate is the same for emotion class *anger*, the accuracy of the best committee trained on Berlin data set is considerably higher (83%) compared to 44% for the best classifier trained on DES data set. For the same

emotion class, the training size is almost three times bigger (127 samples) in the case of Berlin data set than for DES data set (46 samples). The classifiers selected for emotion class *happiness* have similar performance with 71 training samples in the case of Berlin data set and 48 training samples for DES data set.

Table 2: The optimal committees for each emotion class, Berlin data set.

emotion	nf	frames	nr.stages	ac (%)	tpr (%)	fpr (%)
anger	10	1101000001	5	0.83±0.03	0.72±0.16	0.13±0.06
boredom	2	10	58	0.84±0.07	0.49±0.18	0.09±0.09
disgust	10	0100001000	21	0.92±0.05	0.24±0.43	0.00±0.00
anxiety/fear	10	1110000011	86	0.87±0.03	0.38±0.15	0.05±0.04
happiness	10	1111010100	40	0.81±0.06	0.54±0.41	0.14±0.13
sadness	10	1011111101	13	0.91±0.05	0.83±0.06	0.08±0.06

Table 3: The optimal committees for each emotion class, DES data set.

emotion	nf	frames	nr.stages	ac (%)	tpr (%)	fpr (%)
Anger	10	0000100000	11	0.44±0.08	0.72±0.05	0.62±0.08
happy	5	01000	24	0.80±0.01	0.48±0.15	0.13±0.04
neutral	10	0001111000	19	0.83±0.01	0.46±0.28	0.09±0.05
sadness	10	0000011000	6	0.78±0.05	0.35±0.18	0.13±0.12
surprise	10	0011011100	5	0.75±0.11	0.40±0.56	0.18±0.25
targeted	10	1110100110	129	0.97±0.01	0.95±0.07	0.03±0.02

Table 4(for Berlin data set) and table 5(for DES data set) show the influence of the number of frames per utterance on the general recognition results. For each choice of the number of frames, the best classifier is determined against the highest true-positive rate criterion.

Table 4: The dependency of emotion recognition results on the number of frames per utterance for Berlin data set.

nf	ac (%)	tpr (%)	fpr (%)
1	0.85±0.11	0.36±0.63	0.07±0.17
2	0.83±0.31	0.44±0.67	0.10±0.41
3	0.84±0.17	0.46±0.62	0.09±0.23
5	0.84±0.13	0.50±0.63	0.10±0.22
10	0.77±0.33	0.58±0.64	0.20±0.45

The information presented in tables 4 and 5 is independent on the emotion class and so stand for a good comparison criterion. Although the true positive rate tend to be higher for classifiers trained on DES data set, the accuracy rate is still low compared to the accuracy of classifiers trained on Berlin data set. This is associated with the higher false positive rate in the case of DES data and also to the higher classification stability in the case of Berlin data set. One difference should be noted on the analysis methods used for choosing the best classifiers for tables 2, 3 and 4, 5. While for the first the criterion was to choose the classifiers with the best trade-off between hit rate and false positive rate, the last involved the choice for the classifiers with the highest true positive rate. The observation that a 10 frames per utterance is optimal obtained from the tables 2 and 3 can be traced in the performance on the true positive rates from tables 4 and 5.

Table 5: The dependency of emotion recognition results on the number of frames per utterance for DES data set.

nf	ac (%)	tpr (%)	fpr (%)
1	0.61±0.67	0.56±1.12	0.37±1.01
2	0.61±0.67	0.56±1.12	0.37±1.01
3	0.63±0.56	0.56±1.05	0.35±0.87
5	0.63±0.49	0.67±0.90	0.37±0.74
10	0.60±0.55	0.70±1.10	0.42±0.86

CONCLUSIONS AND FUTURE WORK

In the current research we have conducted a set of analysis on different types of utterance segmentation. As a base technique, we used the GentleBoost classifier with a maximum of 200 training stages. The optimal strong classifier has been selected by making use of ROC graphs. The results provided were eventually commented for a better understanding of the underlying phenomena regarding each emotion class. Although the original research includes also separate analysis for male and female voices, the paper presents results only on mixed male-female voices due to the limited amount of space. As a conclusion, we advocate the study of the effect of multi frame speech segmentation as a primary step for the recognition of emotions before actually making a proper choice for a segmentation method and for an efficient recognizer.

REFERENCES

[1] Boersma, P., Weenink, D.: Praat: doing phonetics by computer (Version 4.3.14) [Computer program]. 2005.

[2] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A Database of German Emotional Speech. Proceedings Interspeech, Lissabon, Portugal 2005.

[3] Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C., M., Kazemzadeh, A., Lee, S., Neumann, U., Narayanan, S.: Analysis of emotion recognition using facial expressions, speech and multimodal information ICMI, State College, Pennsylvania 2004.

[4] Chateau, N., Maffiolo, V., Ehrette, T., s'Alessandro, C: Modelling the emotional quality of speech in a telecommunication context. Proceedings of the International Conference on Auditory Display, Kyoto, Japan 2002.

[5] Engberg, I. S., Hansen, A. V.: Documentation of the Danish Emotional Speech Database (DES), Internal AAU report, Center for Person Kommunikation, Denmark, 1996.

[6] Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: A statistical view of boosting. The Annals of Statistics, 38(2) 2000, 337-374.

[7] Kwon, Oh-Wook, Chan, K., Hao, J., Lee, Te-Won: Emotion recognition by speech signals. EUROSPEECH - Geneva (2003) 125–128.

[8] Lee, C., M., Narayanan, S., S.: Toward detecting emotions in spoken dialogs. IEEE Transactions on speech and audion processing 13(2) (2005) 293–303.

[9] 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.

[10] Yu, C., Aoki, P. M., Woodruff, A.: Detecting user engagement in everyday conversations.

ABOUT THE AUTHOR

Dragoş Datcu, PhD Student, Department of Man-Machine Interaction, Delft University of Technology, Phone: +31 15 2783823, E-mail: D.Datcu@ewi.tudelft.nl