Appl Intell (2013) 39:675–691
DOI 10.1007/s10489-013-0449-1

Detecting changing emotions in human speech by machine and humans

C. Natalie van der Wal · Wojtek Kowalczyk

Published online: 15 June 2013

Abstract The goals of this research were: (1) to develop a system that will automatically measure changes in the emotional state of a speaker by analyzing his/her voice, (2) to validate this system with a controlled experiment and (3) to visualize the results to the speaker in 2-d space. Natural (non-acted) human speech of 77 (Dutch) speakers was collected and manually divided into meaningful speech units. Three recordings per speaker were collected, in which he/she was in a positive, neutral and negative state. For each recording, the speakers rated 16 emotional states on a 10-point Likert Scale. The Random Forest algorithm was applied to 207 speech features that were extracted from recordings to qualify (classification) and quantify (regression) the changes in speaker's emotional state. Results showed that predicting the direction of change of emotions and predicting the change of intensity, measured by Mean Squared Error, can be done better than the baseline (the most frequent class label and the mean value of change, respectively). Moreover, it turned out that changes in negative emotions are more predictable than changes in positive emotions. A controlled experiment investigated the difference in human and machine performance on judging the emotional states in one's own voice and that of another. Results showed that humans performed worse than the algorithm in the detection and regression problems. Humans, just like the machine algorithm, were better in detecting changing negative emotions rather than positive ones. Finally, results of applying the Principal Component Analysis (PCA) to our data provided a validation of dimensional emotion theories and they suggest that PCA is a promising technique for visualizing user's emotional state in the envisioned application.

Keywords Affective computing · Vocal expression · Emotion recognition · Speech features · Random forests

1 Introduction

Imagine that your telephone can continuously recognize the emotions you feel, through classifying acoustic features in your voice. The possibilities would be endless. You could use your phone to get insights in your own emotional well-being. Your telephone could be your therapist by listening to your voice and assess your emotional well-being. It could perhaps prevent you from depression by contacting your friends or a professional to help you or give you assignments to feel better. Your phone could also be a social coach, giving you lessons and ratings in how to use your voice to come across positive and enthusiastic instead of lethargic and negative. The current research stems from this vision.

The field of emotion recognition by machines is called affective sensing [15]. Besides the envisioned application indicated above, many other applications of emotion recognition are possible, for example emotion recognition in speech is used in call centers to detect anger in the voice of employees and to give them appropriate feedback [19]. Emotion recognition is useful in real-time conversations with embodied agents in human computer interaction, for example in computer games, but also in (web)applications with virtual therapists. In the multidisciplinary research of emotion,

C.N. van der Wal (✉)
Department of Artificial Intelligence, VU University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands
e-mail: c.n.vander.wal@vu.nl

W. Kowalczyk
Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, 2333 CA Leiden, The Netherlands
e-mail: wojtek@liacs.nl

many definitions of emotions exist. In this chapter, emotion is considered as elicited by a particular stimulus and relatively intense and short lived [8].

Many approaches and techniques are available in emotion recognition research. Emotional expression can be investigated in many different modalities, like gesture, posture, facial expression and speech; e.g. [4, 13, 15, 18, 19]. In [15], the authors claim that affective sensing systems can recognize emotions in human voices and facial expressions better, if semantic features are analyzed as well, besides the standard analysis of sound patterns (like prosody and energy levels), and pattern matching or statistical machine learning techniques in facial expressions. Besides adding semantics to the emotion recognition process, it seems that multimodal emotion recognition gains higher accuracy than uni-modal emotion recognition. For example, in [4] it is shown how emotion recognition through the multiple modalities: facial expressions, body gesture and speech, produces higher accuracy than through any of the single modalities. There are still big steps to be made in research in uni-modal emotion recognition, for example in speech recognition. In [19], the main challenges in automatic emotion recognition from speech are discussed; how to segment audio files, how to extract the relevant features in these speech units and how to develop classifiers on databases with emotional speech with emotional speech. In [1], the authors go deeper into one of these issues: which features in speech are the most important for high accuracy in emotion recognition systems? All of these emotion recognition systems analyze individual uni- or multimodal emotional fragments of human speech, facial expressions or body gestures.

In this paper, the focus is on a single modality, namely speech. One reason for this is that speech is easy to capture/collect. A second reason is that it is easier to process, compared to images/video, which is important for the envisioned application that uses a smart phone processor.

The ultimate goal is to develop a system that will automatically measure changes in the emotional state of a speaker, by analyzing his/her voice, and present the results to the speaker. To achieve this goal the following research questions are addressed: (1) How accurate can the designed intelligent agent predict if a certain emotion in human speech is becoming weaker or stronger? (2) How accurate can it predict how substantial the change is? (3) Do humans perform better or worse than the machine algorithm? (4) How can the emotional states be visualized to the speaker?

The conventional approach of estimation of emotion consists of three steps: (1) Collect human speech as training data, (2) Extract speech features, (3) Use a machine-learning algorithm or statistical analysis to find relations between emotion and speech features, see [12] or [1]. The presented research followed a common approach that is used in pattern classification [5], which involves the following steps: capturing data from sensors (in our case: microphones), data segmentation (in our case: manual splitting of recordings into meaningful pieces and labeling them), feature extraction (converting each recording into a vector of features of fixed length) and developing a classification or regression model (we used Random Forests). Random Forests were used to address a classification problem (out of 2 recordings, which one has the highest emotional value?) and a regression problem (given 2 recordings, estimate the difference in each emotional state).

The paper is organized as follows: in Sect. 2 the main steps of our approach to the problem are explained: the data capturing, data segmentation, feature extraction and data modeling. Section 3 describes the results of the machine learning algorithm. Section 4 describes the validation experiment: human performance on the same task and shows the results. Section 5 addresses the problem of visualizing the multidimensional emotional data in 2-d space. Finally, Sect. 6 interprets the results and in Sect. 7, challenges, possible extensions and applications of this research are discussed.

2 Method

In the presented research we followed the standard steps: data collection, data preprocessing, feature extraction, preparation of training sets, and development of classification and regression models.

In the current experiments, Random Forests were used [2, 10] for their superior accuracy, good generalization properties and ease of training. An alternative method to Random Forests could be Support Vector Machines (SVM) with some well chosen kernel functions [17]. However, algorithms for training SVM are very sensitive to the choice of learning parameters and are computationally more demanding than Random Forests.

2.1 Data collection

The speech samples used in this research were collected from 77 participants (Dutch university students and employees, 34 women, mean age: 26.77 years) during a laboratory study in which the participants were induced with positive, neutral or negative emotions. Participants were tested on the effect of emotion on different cognitive capabilities, by watching a 5-minute movie that induced a certain emotion and afterwards performing a verbal test that measured their creativity, attention or action urges. The speech of the participants was recorded with a ZOOM H1 voice recorder, through a clip-on microphone, with the automatic gain control on. Participants were informed that their speech

was recorded during the experiment, so that the researcher could write out their answers afterwards. Participants had no clue that their speech was recorded for the purpose of emotion recognition by a machine. In this way, naturally occurring speech was collected. This is beneficial for this study, because it does not have the pitfalls of acted emotional speech. Acted speech elicits how emotions should be portrayed, not necessarily how they were portrayed. Also, acted emotions do not emerge in the body and mind of the individual as naturally occurring emotions do. Another advantage of this data collection method is that the participants own ratings of their emotions were collected. Each of the 77 participants recorded, in 3 independent sessions, a short message (containing their name, age and the city of residence), while being in one of the 3 emotional states: "neutral" (entering the lab), "positive" (after watching a short "positive" movie), and "negative" (after watching a short "negative" movie). Next, each subject rated his/her emotional states on an Emotion Report Form containing the following 16 emotions: sadness, fear, shame, contentment, guilt, happiness, disgust, despair, positivity, enjoyment, irritation, hope, anger, pride, negativity, anxiety. Ratings were made on a 10-point Likert scale (1 = none, 10 = a great deal).

2.2 Data segmentation

The speech units were cut manually, starting exactly at the first phoneme spoken by the subject and ending directly after the last phoneme was spoken; periods of silence at the beginning and at the end of each recording were manually removed. As in realistic speech databases, the word itself is normally not the optimal emotion unit to be processed, the speech units collected in this research consist of several words. Each audio segmentation represents a unit that is representative for a positive or negative emotion. The name-age-city units establish a syntactically/semantically meaningful unit, in which we assume the felt emotion can be expressed easily by the participant because he/she can identify him/herself with the semantic meaning of the words (and even with the sounds of the letters/words). The units differ in length for each subject, varying from 2 to 8 seconds. Each speech recording, originally stored in a WAV format (stereo, 44.100 Hz frequency sampling), was converted into a single vector of numbers (signals from the two channels were combined into one: $s_{mono} = (s_{left} + s_{right})/2$) and stored in Matlab.

2.3 Feature extraction

In order to apply any classification or regression algorithm to our recordings, each recording had to be converted to a vector of fixed length of sound features. It was difficult to say beforehand, which features would be most successful. Therefore almost all features that could be found in the available literature on analyzing affective speech: [1, 3, 6–11] were used. These features were calculated in two steps:

(1) Each recording was split into a sequence of short segments (10 milliseconds long), and a number of procedures were applied to each segment to calculate, among others, features like: fundamental frequency, energy, formants, Cepstral coefficients.
(2) The values of the computed features were aggregated over the duration of the whole recording. The aggregates included mean, median, standard deviation, skewness.

Table 1 provides a complete overview of speech properties and aggregates that were used in the current experiments.

2.4 Training sets

After calculating all features for each recording, the final training set was prepared for developing classification

Table 1 All 207 speech features used by the classification algorithm

23 speech properties	9 aggregates for each speech property
F0 (Fundamental Frequency)	Mean
I (Sound Intensity, measured on the logarithmic scale)	Median
E (Sound Energy, measured in sound units)	Standard deviation
F1, F2, F3 (first 3 formants)	Skewness
B1, B2, B3 (bandwidth of first 3 formants)	Kurtosis
MFCC0-MFCC12 (13 Mel Frequency Cepstral Coefficients)	Q1 (mean of the smallest 20 % of values)
SR (speech restarts)	Q5 (mean of the biggest 80 % of values)
	Shimmer (period to period variability)
	Rise (percentage of times next value is bigger than the previous one)

and regression models. The following 2 problems were addressed:

Classification: Develop, for each of the 16 emotional states, a classification procedure, which, when applied to two recordings S1 and S2 will determine if the emotional state in S2 is "more present" than in S1 (has higher intensity value).

Regression: Develop, for each of the 16 emotional states, a regression procedure, which, when applied to two recordings S1 and S2 will estimate the difference of the intensity of the emotional state between S1 and S2.

Additionally, it was assumed that both recordings, S1 and S2, are coming from the same speaker and the input for the classification and regression procedure consists of the differences between feature vectors of S1 and S2, and not the original values. In other words, the change in emotional state has to be predicted from the change in feature vectors. The main reason for this assumption was caused by the scarcity of data: for each subject there were only 3 recordings and (big) interpersonal differences in speech features would dominate the subtle changes that reflect emotional states.

The training set was constructed as follows. Let S1, S2, S3 denote 3 recordings of the same subject. This triplet leads to 3 input vectors: S1–S2, S1–S3, S2–S3, and the corresponding output values: E1–E2, E1–E3, E2–E3 (in the regression task), or $+1$ or -1 (in the classification task), depending on the sign of the difference (cases where the difference was 0 were ignored). In total, the training set used for developing our regression and classification models had 225 records.

2.5 Modeling data with Random Forests

The concept of Random Forests was introduced by Leo Breiman in 2001 [2], and since then it became one of the most prominent techniques for solving classification and regression problems [5]. The key idea behind this technique is a construction of many (hundreds) of de-correlated classification or regression trees and then aggregating their predictions. This leads to models with very good accuracy. The construction of a random forest of K trees for a set on N records is as follows:

> Repeat steps (1) and (2) K times to develop K trees:
> (1) draw a random sample of N records (with replacement) from the available data,
> (2) develop a classification or regression tree for the data sample in the following way: (a) whenever a splitting attribute has to be chosen, consider all possible attributes and select at random one of the top L best attributes, (b) whenever a node covers M (or less) records don't split it anymore.

When a Random Forest is applied to new data, outputs of all trees are either averaged (in case of regression), or the most frequent output label is chosen as a result (in case of classification). The Random Forest procedure involves 3 parameters: the number of trees to be developed, K, the number of best splitting attributes L, and the limit on the leaf size, M. In practice, the choice of K (the number of trees) is not critical: it has to be sufficiently big and further increase of this value has no impact on the model accuracy. In our case, K was set to $K = 200$. The optimal value of L was determined experimentally, by trying values $L = 25, 50, 75, 100$; the best results were achieved for $L = 75$. Finally, the choice of the minimal number of records in a leaf, M, was most difficult: it strongly depends on the emotional state that we wanted to model (different states needed different values of M). To choose this value, the following heuristic was used. For each emotion, 6 values of M $(1, 2, \ldots, 6)$ were tried, choosing the one that was best (had smallest average error) in the interval of 150–200 trees.

To avoid data overfitting, out-of-bag error estimates were used, as described in [5]. These estimates are computed as follows. Each tree from a random forest is trained on a sample of data (a "bag"). Because data is sampled with replacement, some records are not used in the training (they are "out-of-bag") so they can be used as a test sample to estimate the accuracy of the trained tree. Clearly, each tree is trained and tested on a different sample, therefore the average accuracy of all trees, measured on "out-of-bag" samples, gives a very reliable estimate of the true accuracy of the random forest.

The process of training random forests is time consuming. Its time complexity is proportional to the number of trees, the number of features and the number of records in the training dataset, and it also depends on the leaf size limit. In our situation, training a random forest for a single emotion took several minutes on a simple laptop computer (with the Intel P8400, 2.26 GHz, processor). However, applying a random forest to a vector of features requires only a single scan of all the trees and it can be accomplished in a fraction of a second. The process of calculating speech features, which is based on the Discrete Fast Fourier Transform algorithm, is very fast and can be executed almost simultaneously with capturing the sound signal. Thus, although the process of training random forests takes minutes, analyzing speech and predicting the corresponding intensity of emotions can be done almost in real time.

3 Results

The results of the experiments are summarized in Tables 2 and 3. In case of classification models, the error was measured by the percentage of misclassified cases. This error was then compared to the baseline error: the error made by

Table 2 Results of the classification problem. Error is measured by the ratio of misclassified records (misclassification rate)

Emotion	Baseline error	Model error	Relative error reduction
Sadness	0.4024	0.3491	13.24 %
Fear	0.3924	0.2975	24.19 %
Shame	0.4397	0.2414	45.10 %
Content	0.3109	0.3057	1.67 %
Guilt	0.3578	0.3853	−7.69 %
Happiness	0.4341	0.4439	−2.25 %
Disgust	0.2628	0.2564	2.44 %
Despair	0.3889	0.3175	18.37 %
Positivity	0.3418	0.3316	2.99 %
Enjoyment	0.399	0.3695	7.41 %
Irritation	0.3113	0.2914	6.38 %
Hope	0.3242	0.3626	−11.86 %
Anger	0.3514	0.3514	0.00 %
Pride	0.2442	0.2558	−4.76 %
Negativity	0.3714	0.3429	7.69 %
Anxiety	0.3168	0.2857	9.80 %

Table 3 Results of the regression problem. The Mean Squared Error measure is used (MSE)

Emotion	Baseline error	Model error	Relative error reduction
Sadness	9.1392	8.5159	6.82 %
Fear	12.1157	10.2778	15.17 %
Shame	2.7825	2.4361	12.45 %
Content	12.7282	11.6326	8.61 %
Guilt	4.2478	4.0052	5.71 %
Happiness	15.7575	14.2149	9.79 %
Disgust	16.5909	14.4593	12.85 %
Despair	7.5872	6.7577	10.93 %
Positivity	15.3961	14.0813	8.54 %
Enjoyment	17.1388	16.2514	5.18 %
Irritation	8.2867	7.2742	12.22 %
Hope	10.0526	9.4538	5.96 %
Anger	10.4548	9.5458	8.69 %
Pride	7.8706	7.4964	4.75 %
Negativity	13.6128	11.5905	14.86 %
Anxiety	15.0084	13.1273	12.53 %

the base classifier that always predicts the most frequent category. In case of regression models, the error measure was the Mean Squared Error, MSE. The baseline model was defined as a constant function equal to the mean value of the predicted variable.

Table 4 Most important attributes in classification and regression models of detecting changing emotions and their intensity

Most important attributes	Frequency
M2std: the standard deviation of the 2nd cepstral coefficient	38 times
M6mean: the mean value of the 6th cepstral coefficient	16 times
PU: the period to period variability(shimmer) of speech energy	12 times
I: the speech intensity	12 times
M6: the 6th cepstral coefficient	12 times

Finally, the Relative Error Reduction (RER) was calculated which is defined as the ratio (BaselineError-ModelError)/BaselineError.

3.1 Most important features

The Random Forest algorithm provides a powerful mechanism for measuring the importance of attributes that are used in the modeling process [5]. To measure the importance of an attribute, its values are permuted and the original accuracy of the model (measured on out-of-bag samples) is compared to the accuracy of the model on the modified data (on the same out-of-bag samples). The observed difference between both accuracies is strongly related to the importance of the attribute: the bigger the difference the more important the attribute.

With the help of this method, the importance of each attribute was found and calculated, for each emotion. Due to lack of space, only the most frequent attributes that were used by Random Forests are listed. More precisely, for each model, the five most important attributes were listed, all 32 lists (16 for classification and 16 for regression) were concatenated and frequencies of the attributes on the list were computed. The most frequent attributes are listed in Table 4. The identification of the most informative features has 3 objectives: (1) verification of our findings with the existing literature, (2) simplification of the implementation of the automated system for monitoring emotions, (3) better understanding of the emotion-speech relation.

4 Validation experiment

In this section, the results of the controlled experiment, in which humans perform the same task as the machine algorithm in the previous section, are presented. The following research questions will be examined:

(1) Are humans better or worse than the machine algorithm in the classification of emotions? (out of 2 recordings, which one has the highest emotional value?).

(2) Are humans better or worse than the machine algorithm in the regression problem? (Given 2 recordings, estimate the change of each emotional state).
(3) Are humans better in detecting changing emotions in negative than positive emotions, like the machine algorithm?
(4) Are humans better in predicting changing emotions in their own voice than somebody else's voice?
(5) Are humans better in detecting changing emotions when they can also use semantic cues in the audio files, besides the acoustic cues?
(6) Are humans better in detecting changing emotions when they learn how the speaker rated his own feelings than when they did not learn this?

For questions 1 and 2, the machine algorithm is expected to perform better than humans, because humans cannot find semantic cues in these recordings. Also, the acoustic emotional cues are difficult for humans to hear, because the recordings sound like the speakers are repeating a memorized sentence, a bit monotonic/non-emotional. For question 3, there is no expectation beforehand. Regarding question 4, it is expected that humans perform better in emotion recognition in their own voice, because we tend to know our own voice better than a stranger's voice. For question 5, it is expected that humans perform better when they can also use semantic cues in the recordings, like "I would like to yell and scream" or "I would like to jump for joy". Finally, regarding question 6, it is expected that humans perform better when they learn by seeing the speaker's own emotional ratings for one recording, than when they do not.

4.1 Method

4.1.1 Participants

Experiment 1. The total number of participants (a subset of the participants from experiment 1) was 12 university students and employees (7 women, average age 31.9 years).

Experiment 2, 3, 4 and 5. The total number of participants was 14 university students and employees (7 women, average age 31.6 years).

Participants for all experiments were randomly selected from the subject pool of 80 participants that entered the original experiment 5 months earlier in which their voices were recorded. Each participant received a chocolate bar as reward.

4.1.2 Materials

Written. Emotional experiences were assessed using an Emotion Report Form, based on the Geneva Appraisal Questionnaire (GAQ, 2002) and on the emotion report form used by Frederickson and colleagues [7]. Participants rated how much they thought the speaker was feeling each of the following 16 emotions: sadness, fear, shame, contentment, guilt, happiness, disgust, despair, positivity, enjoyment, irritation, hope, anger, pride, negativity, anxiety. Ratings were made on a 10-point Likert scale (1 = none, 10 = a great deal).

Auditory. 168 audio files served as the experimental manipulation in this research.

In the original experiment, 5 months earlier, speech was collected from the participants, as explained in Sect. 2.1. Four different types of audio files were manually divided. The first type, in which the participant stated his/her name, age and city of residence, was used as the training and test sets of the emotion recognition system, as described in the previous section. In the current validation experiment three other types of audio files were used as well. The reason for this was to be able to compare human performance on audio files with and without semantic information. The four types of audio files are explained in Table 5. Especially in the "TAR" and "CR" files, the participants could find semantic cues to how the speaker feels at that moment, either positive or negative. (For example, if a person felt angry, he/she would state he/she would like to fight and yell, and

Table 5 The 4 types of audio files used in the validation experiment

Type of audio file	Explanation
"NAC" (Name-Age-City)	The speaker states his name, age and city of residence
"TAR" (Thought Action Repertoire)	The speaker states all the things he/she would like to do right now (based on his/her current emotions). For example: "I would like to go surfing, snowboarding, uhm... go out with friends, ... uhm and drink a cup of coffee"
"MEM" (Memory)	The speaker reads out the first question of the memory task and the answer. For example: "What color is the tie of Mr. Bean? The answer is purple/green"
"CR" (Creativity)	The speaker states all things he can think of, one can do with a newspaper. For example "Uhm..., you can fold it into a boat, you can kill a fly with it, uhm..., you can recycle it, uhm..."

Table 6 The 3 conditions of the experiment

Condition	Explanation
"Self"	Participants listened to 4 speech recordings of themselves, recorded in the laboratory study they entered 5 months ago.
"Other-without-learning"	Participants listened to 4 speech recordings of another person, recorded during the same laboratory study they entered 5 months ago.
"Other-with-learning"	Participants listened to 4 speech recordings of another person, recorded during the same laboratory study they entered, 5 months ago. During the first recording, the participant could study the Emotion Report Form filled in by the speaker (to learn to detect the right emotional state in the voice of the speaker).

when a person was happy, he/she would state he/she would like to dance and smile.)

Each participant listened to a total of 12 speech recordings: 4 recordings of themselves and 8 recordings of two other persons. There were three conditions in this experiment: "Self", "Other-without-learning" and "Other-with-learning", explained in Table 6. Each participant performed each condition. For all participants, the orders of the three conditions and of the four audio files in each condition were randomized, to control for order effects.

4.1.3 Dependent measure

The ratings of humans on changing emotions in the speaker were made on the Emotion Report Form, on which 16 emotions could be rated: sadness, fear, shame, contentment, guilt, happiness, disgust, despair, positivity, enjoyment, irritation, hope, anger, pride, negativity, anxiety. Ratings were made on a 10-point Likert scale (1 = none, 10 = a great deal). In the original experiment, the speakers would have filled in the Emotion Report Form as well. From these ratings, different calculations were made to answer the different research questions. For research questions 1, 2 and 3 (to compare human and machine performance in the same way) the participants' ratings of subsequent audio files of the same speaker were compared in the same way as recordings S1–S2, S2–S3 and S3–S1 were compared in the previous section. The only difference is that in this experiment the orders of the files were preset, namely 111 different orders in Experiment 2 and 20 different orders in Experiment 1. The errors were calculated in the same way as the classification and regression problems of the machine algorithm. For the classification problem, the error was measured by the percentage of misclassified cases. This error was then compared to the baseline error: the error made by the base classifier that always predicts the most frequent category. In case of regression models, the error measure was the Mean Squared Error, MSE. The baseline model was defined as a constant function equal to the mean value of the predicted variable. For both classification and regression problems, the Relative Error Reduction (RER) was defined as the ratio (BaselineError-ModelError)/BaselineError. To answer research questions 4 (Do humans detect changing emotions better in their own voice than another person's voice), the human performances of condition "Self" was compared to conditions "Other-without-learning" and "Other-with-learning" together in Experiment 4. To answer research question 5 (Do humans perform better when using semantic cues + acoustic cues compared to only acoustic cues?), the human performances on the "NAC" audio files were compared to the performances on the "MEM", "TAR" and "CR" audio files together in Experiment 5. Finally, to answer research question 6 (Do humans perform better in the 'learning' condition than the 'non-learning' condition?), the human performances on the condition "Other-with-learning" was compared to "Other-without-learning" in Experiment 6. For research question 4, 5, and 6, the average ratings on each emotional state, per stated condition, were compared.

4.1.4 Procedure

Participants were tested individually. On arrival, they were seated behind a desk and provided their informed consent. A laptop with headphones was located in front of the participant, next to the written material: the Emotion Report Forms and the instruction of the experiment. After reading the instructions, the participant would use the headphone to listen to 12 recordings in the order instructed to him. Each participant would listen to 4 recordings of himself, and 8 recordings of 2 other persons (4 each). After listening to each recording, the participant reported the emotional states of the speaker on an Emotion Report Form. In one case, the participant would read the Emotion Report Form filled in by the speaker to be able to learn which emotional values belong to the acoustic (and semantic) features in the speaker's voice. After filling in the last Emotion Report Form, the participant was thanked for his/her work and rewarded with a chocolate bar.

Table 7 Results of Experiment 1: the humans on the classification problem on only "NAC" files, like the machine algorithm (only 40 recordings of 20 speakers; 20 pairs)

Emotion	Human performance			Machine performance			p-Value
	N	N_{correct}	Accuracy	N	N_{Correct}	Accuracy	
Sadness	19	7	37 %	169	110	65 %	0.016
Fear	19	8	42 %	158	111	70 %	0.014
Shame	11	8	73 %	116	88	76 %	0.817
Content	20	7	35 %	193	134	69 %	0.002
Guilt	10	3	30 %	109	67	61 %	0.053
Happiness	20	8	40 %	205	114	56 %	0.181
Disgust	18	2	11 %	156	116	74 %	0.000
Despair	14	6	43 %	126	86	68 %	0.058
Positivity	20	11	55 %	196	131	67 %	0.288
Enjoyment	20	9	45 %	203	128	63 %	0.114
Irritation	18	5	28 %	151	107	71 %	0.000
Hope	18	8	44 %	182	116	64 %	0.108
Anger	17	3	18 %	148	96	65 %	0.000
Pride	16	8	50 %	172	128	74 %	0.037
Negativity	19	9	47 %	175	115	66 %	0.114
Anxiety	20	6	30 %	161	115	71 %	0.000

4.2 Results

In this section, the results of the controlled experiments are presented.

4.2.1 Human vs. machine performance

In this subsection, the following research questions are answered:

(1) Are humans better or worse than the machine algorithm in the classification of emotions? (Out of 2 recordings, which one has the highest emotional value?).
(2) Are humans better or worse than the machine algorithm in the regression problem? (Given 2 recordings, estimate the change of each emotional state).
(3) Are humans better in detecting changing emotions in negative than positive emotions, like the machine algorithm?

To make a direct comparison between humans and the machine algorithm, the performance of the humans on only "NAC" files is shown in Tables 7 and 8. The a priori hypothesis was that the machine learning algorithm would outperform the humans because the NAC files contain no semantic cues for the humans and the voices sound a bit 'monotone' or 'emotionless'. The significance of results presented in Table 8 are analyzed using a one-tailed unpaired t-test based on the a priori hypothesis. The significance of results presented in Table 7 were analyzed with a χ^2 test because the predicted variable is nominal.

Answering question 1: compared to the machine algorithm, humans did not outperform the machine algorithm on any emotion in the classification problem of the "NAC" files. Actually, the machine algorithm outperformed the humans on all emotions, from which the following have significant performance differences: sadness, fear, content, disgust, irritation, anger, pride, anxiety. No significant difference was found for: shame, happiness, positivity, enjoyment, hope and negativity. A trend was found for guilt and despair, where the machine algorithm performed slightly better in detecting changes in these emotions than humans.

Looking at the regression problem, question 2, in comparison with the machine algorithm, humans performed worse on all emotions. Significant differences were found for almost all emotions, namely: sadness, fear, content, happiness, disgust, despair, positivity, enjoyment, irritation and anxiety. For the other emotions, shame, hope anger, and pride, a trend was found. Guilt is the only emotion for which no significant difference was found. These results indicate that the machine algorithm is better in predicting the size of change in intensity of an emotion than humans.

Answering question 3: the results do not indicate that humans are better in detecting changes in negative emotions than positive emotions, based on the accuracies.

Note that the results on the "NAC" files only (Tables 7 and 8) have to be taken with caution because of the modest number of recording pairs in the experimental setup for humans (only 40 recordings of 20 speakers; 20 pairs), while the training and test sets of the machine algorithm could use a total of 225 recordings of 75 different speakers; 225 pairs. For the classification problem, two classes were chosen: "increase" or "decrease" of intensity of an emotion. Therefore, for each emotion, only pairs of recordings were considered,

Table 8 Results of Experiment 1: the humans on the regression problem on only "NAC" files, like the machine algorithm (40 recordings, 20 pairs of 20 different speakers)

Emotion	Human performance			Machine performance			*p*-Value
	N	Mean absolute difference	Standard deviation	*N*	Mean absolute difference	Standard deviation	
Sadness	20	3.95	2.65	225	2.2103	1.91	0.000
Fear	20	4.125	2.19	225	2.4451	2.08	0.002
Shame	20	1.4	1.05	225	1.0665	1.14	0.094
Content	20	4.65	3.27	225	2.6842	2.11	0.000
Guilt	20	1.7	2.08	225	1.2671	1.55	0.187
Happiness	20	5.15	2.91	225	3.0792	2.18	0.003
Disgust	20	5.15	2.35	225	3.025	2.31	0.001
Despair	20	3.2	2.67	225	1.8274	1.85	0.002
Positivity	20	4.7	2.75	225	3.0492	2.19	0.008
Enjoyment	20	4.65	3.08	225	3.3414	2.26	0.009
Irritation	20	3.8	2.02	225	2.0603	1.74	0.001
Hope	20	3.15	2.11	225	2.4571	1.85	0.085
Anger	20	3.25	2.34	225	2.3607	2.00	0.057
Pride	20	2.9	2.31	225	2.1387	1.71	0.083
Negativity	20	3.95	3.30	225	2.7096	2.07	0.008
Anxiety	20	5.4	2.96	225	2.7788	2.33	0.001

Table 9 Results of Experiment 2: the humans on the classification problem on all four types of files (168 recordings of 42 speakers; maximum of 111 pairs for the humans). Error is measured by the ratio of misclassified records (misclassification rate)

Emotion	Human performance			Machine performance			*p*-Value
	N	N_{correct}	Accuracy	*N*	N_{Correct}	Accuracy	
Sadness	70	27	39 %	169	110	65 %	0.000
Fear	70	31	44%	158	111	70%	0.000
Shame	41	24	59 %	116	88	76 %	0.035
Content	78	44	56 %	193	134	69 %	0.041
Guilt	34	11	32 %	109	67	61 %	0.003
Happiness	78	41	53 %	205	114	56 %	0.646
Disgust	72	21	29 %	156	116	74 %	0.000
Despair	53	23	43 %	126	86	68 %	0.002
Positivity	78	47	60 %	196	131	67 %	0.303
Enjoyment	78	48	62 %	203	128	63 %	0.814
Irritation	65	22	34 %	151	107	71 %	0.000
Hope	71	36	51 %	182	116	64 %	0.057
Anger	60	20	33 %	148	96	65 %	0.000
Pride	61	23	38 %	172	128	74 %	0.000
Negativity	72	33	46 %	175	115	66 %	0.004
Anxiety	70	29	41 %	161	115	71 %	0.000

that had different emotional intensity. For example, for the humans, there were only 11 pairs of recordings with different intensity of "shame"; the remaining 9 pairs had the same intensity of this emotion and were not taken into account, see Table 7. To make a better comparison between the machine and the humans, based on the number of recordings, experiment 2 was performed, in which the number of recording pairs of the humans are comparable to that of the machine algorithm.

Tables 9 and 10 show the results of human performance on the classification and regression problems in Experiment 2. These results are calculated over all types of files that humans listened to. The machine algorithm was only trained and tested on "NAC" files, which seems to give humans a benefit of having semantic cues on top of the acoustic cues, especially in the "CR" and "TAR" files. The a priori hypothesis was that the machine learning algorithm would perform differently than the humans, but because the hu-

Table 10 Results of Experiment 2: the humans on the classification problem on all four types of files (168 recordings of 42 speakers; maximum of 111 pairs for the humans)

Emotion	Human performance			Machine performance			p-Value
	N	N_{correct}	Accuracy	N	N_{Correct}	Accuracy	
Sadness	111	2.6216	2.40	225	2.2103	1.91	0.09
Fear	111	2.8378	2.59	225	2.4451	2.08	0.135
Shame	111	1.4775	1.35	225	1.0665	1.14	0.004
Content	111	3.2162	2.77	225	2.6842	2.11	0.052
Guilt	111	1.3784	1.83	225	1.2671	1.55	0.582
Happiness	111	3.6126	2.82	225	3.0792	2.18	0.057
Disgust	111	3.2523	2.72	225	3.0250	2.31	0.424
Despair	111	2.0721	2.31	225	1.8274	1.85	0.296
Positivity	111	3.3243	2.62	225	3.0492	2.19	0.341
Enjoyment	111	3.1622	2.90	225	3.3414	2.26	0.535
Irritation	111	2.4324	1.98	225	2.0603	1.74	0.08
Hope	111	2.7297	2.39	225	2.4571	1.85	0.251
Anger	111	2.0991	2.19	225	2.3607	2.00	0.291
Pride	111	2.5045	2.12	225	2.1387	1.71	0.09
Negativity	111	3.0721	2.64	225	2.7096	2.07	0.17
Anxiety	111	3.3784	2.99	225	2.7788	2.33	0.045

mans now had an advantage compared to Experiment 1 (now they could also use semantic cues), there was no hypothesis formed a priori about who would perform better, the humans or the machine algorithm. The results in Table 10 were therefore analyzed using two-tailed unpaired t-tests. The results in Table 9 were analyzed with a χ^2 test, because the data acquired are nominal.

Answering questions 1 and 2: even though the humans had an advantage compared to the machine learning algorithm, the humans performed worse than the machine algorithm on both the classification and regression problems. First of all, on the classification problem, compared to the human performance, the machine algorithm detected changes in sadness, fear, shame, content, guilt, disgust, despair, irritation, anger, pride, negativity and anxiety significantly better. For hope a trend was found; $p = 0.251$. No significant differences were found for happiness, positivity and enjoyment. The overall performance of the machine algorithm on the classification problem is therefore better than the human performance: significantly better on 12 out of 16 emotions. When looking at the regression problem, humans in comparison with machine, the machine algorithm performed significantly better on sadness, shame and anxiety. For content, happiness, irritation and pride trends were found; $p = 0.052$, $p = 0.057$, $p = 0.08$ and $p = 0.09$, respectively. On all other emotions, the performances did not differ significantly.

Answering question 3: based on the accuracies, there seems to be no clear difference between detecting positive versus negative emotions for humans.

4.2.2 More characteristics of human performance

In this subsection, the answers to the following research questions are given:

(4) Are humans better in predicting changing emotions in their own voice than somebody else's voice?
(5) Are humans better in detecting changing emotions when they can also use semantic cues in the audio files, besides the acoustic cues?
(6) Are humans better in detecting changing emotions when they learn how the speaker rated his own feelings than when they did not learn this?

Below, in Table 11, the results of Experiment 3 are shown. The average differences of the ratings of the participants on the speaker's felt emotions, based on their recordings are shown. For all types of files, for all emotions except pride, the participants were better in rating the emotional intensities of their own voice than that of others. When only looking at the "NAC" files, on all emotions except disgust and pride, the participants were better in rating the emotional intensities of their own voice than that of others. When only looking at the non-"NAC" files, on all emotions except enjoyment, hope and pride, the participants were better in rating the emotional intensities of their own voice than that of others. When looking at the averages over all emotions, the participants were always better in rating the emotional intensities of their own voice than that of others. These results indicate that it is easier to detect the intensities of emotions in one's own voice than that of a stranger.

Table 11 Results of Experiment 3: average absolute difference in emotional ratings of one's own voice and of another person's voice

Emotion	All type of files		Only "NAC" files		Non-"NAC" files	
	Average absolute difference own voice	Average absolute difference other person's voice	Average absolute difference own voice	Average absolute difference other person's voice	Average absolute difference own voice	Average absolute difference other person's voice
Sadness	1.64**	2.52**	1.82	2.59	1.46**	2.43**
Fear	1.93*	2.56*	2.21	2.38	1.64**	2.75**
Shame	1.36**	2.02**	1.36	1.90	1.36**	2.15**
Content	2.32	2.43	2.61	2.63	2.04	2.22
Guilt	1.00***	2.19***	0.96**	1.84**	1.04***	2.57***
Happiness	2.50	2.75	2.61	3.02	2.39	2.46
Disgust	2.63	2.64	3.00	2.69	2.25	2.59
Despair	1.50*	2.14*	1.36	2.10	1.64	2.20
Positivity	2.07*	2.65*	2.07	2.75	2.07	2.54
Enjoyment	2.32	2.60	2.29	3.00	2.36	2.15
Irritation	2.39	2.77	2.39	2.80	2.39	2.74
Hope	2.16	2.20	2.04	2.22	2.29	2.17
Anger	1.70	2.06	2.00	2.12	1.39	2.00
Pride	2.75	2.47	3.07	2.67	2.43	2.26
Negativity	1.93**	2.72**	1.86*	2.88*	2.00	2.54
Anxiety	2.20	2.77	2.50	2.55	1.89**	3.02**
Average over all emotions	2.02***	2.47***	2.13**	2.51**	1.92***	2.42***

* $p < .1$, ** $p < .05$, *** $p < .001$

In Table 12, the results of Experiment 4 are shown. The average differences of the ratings of the participants on the speaker's felt emotions, based on their recordings are shown. The performances on the "NAC" files are compared to that of the "MEM", "TAR" and "CR" files together. Differences could indicate that semantic cues are the reason for a better performance on the non-"NAC" files, since there were no semantic cues in the "NAC" files. The results show that the participants performed better in rating sadness, content, happiness, disgust, positivity, enjoyment, irritation, anger, pride and negativity in the non-"NAC files conditions, but worse in rating fear, shame, guilt, despair, hope and anxiety.

When looking at the averages over all emotions, the participants were marginally better in rating the emotional intensities of the non-"NAC" files than the "NAC" files. These results do not confirm the expectation that humans can rate the intensity of recordings that include acoustic and semantic cues better than recordings with acoustic cues alone. Below, in Table 13, the results of Experiment 5 are shown. The table contains average differences of the ratings of the participants on the speaker's felt emotions, based on their recordings. The performances under the learning condition are compared to the condition without learning. Differences could indicate that learning improves human performance, which was the expected finding. The results show that the participants performed better in rating fear, content, happi-

Table 12 Results of Experiment 4: average absolute differences in emotional ratings of "NAC" files versus "CR", "TAR" and "MEM" files together

Emotion	Average absolute difference "NAC" files	Average absolute difference "MEM"+ "CR" + "TAR" files
Sadness	2.32	2.07
Fear	2.32	2.33
Shame	1.71	1.85
Content	2.62	2.15
Guilt	1.53	1.99
Happiness	2.87	2.43
Disgust	2.80	2.46
Despair	1.84	1.99
Positivity	2.51	2.36
Enjoyment	2.75	2.23
Irritation	2.66	2.61
Hope	2.15	2.22
Anger	2.08	1.77
Pride	2.81	2.32
Negativity	2.52	2.34
Anxiety	2.53	2.59
Average over all emotions*	2.38	2.23

* $p < .1$

Table 13 Results of Experiment 5: average absolute differences in emotional ratings of the condition with learning versus without learning

Emotion	Average absolute difference with learning	Average absolute difference without learning
Sadness	2.23	2.17
Fear	2.13	2.38
Shame	1.98	1.71
Content	2.28	2.44
Guilt	1.80	1.71
Happiness	2.60	2.68
Disgust	2.35	2.76
Despair	1.83	1.97
Positivity	2.60	2.37
Enjoyment	2.20	2.57
Irritation	2.55	2.68
Hope*	1.63	2.37
Anger	1.85	1.99
Pride**	1.50	2.94
Negativity	2.48	2.40
Anxiety	2.43	2.59
Average over all emotions*	2.15	2.36

*$p < .05$, **$p < 0.001$

ness, disgust, enjoyment, irritation, hope, anger, pride and anxiety, but not in rating: sadness, shame, guilt, positivity and negativity. When looking at the averages over all emotions, the participants were marginally better in rating the emotional intensities under the learning condition than under the non-learning condition. These results indicate that overall, humans can rate the emotional intensity of recordings when they have seen one of the speaker's own ratings with one of his recording marginally better than when not.

5 Visualization

In this section, the task to reduce the dimensionality of the emotional data (originally 16 dimensions) and to visualize it is addressed. At the beginning of this chapter, an intelligent agent was envisioned that is built into a phone and who can measure the emotion in the voice of its user. The previous sections addressed a machine algorithm that can endow an intelligent agent with emotion recognition from speech. The next goal is to design a cognitive model for this agent, so that it can visualize multidimensional data collected from its user in 2-*d* space. How exactly can an intelligent agent give an objective insight into your emotions? By showing you your own face, as you are looking into a mirror? That could fail because we often don't really see ourselves properly in the mirror. Think of people with an eating disorder that perceive themselves as being overweight in the mirror, while in fact they are underweight. The same can happen with emotions: you could perceive your own expressed emotion stronger or weaker than it really is. Therefore another visualization technique is necessary. Here, it was chosen to use a dimensionality reduction technique which can map multidimensional data into two dimensions, while preserving as much information as possible.

There are many methods for dimensionality reduction, for example: Principal Component Analysis (PCA), Multidimensional Scaling (MDS), Locally Linear Embedding (LLE) and Self-Organizing Maps (SOM) [14]. For this research PCA analysis was chosen, for several reasons. First, PCA is computationally very cheap and can be performed on a smart phone. In contrast to PCA, MDS and LLE would require the whole algorithm to be rerun every single time a new speech unit is analyzed. Second, PCA requires no tuning and is fully deterministic. Furthermore, PCA is a well-established standard in psychometrics and provides closed formulas to compute coordinates of data in new dimensions. MDS and LLE are more visualization techniques; there is no formula that describes the mapping. SOM is also a visualization technique; it is not an analytical technique that measures the structure in the data.

5.1 Dimensional theory of emotion

There are different theories of emotions, amongst others: dimensional models of emotions and the theory of basic emotions [6, 16]. Dimensional models of emotions assume that emotions are dependent upon each other, whereas, the basic emotions theory assumes that emotions are independent of each other. According to the basic emotions theory, emotions can be labeled as anger, disgust, fear, happiness, sadness and surprise, see [6]. These labels do not give many options for the visualization of emotions. For the purpose of visualizing emotions to the user of our envisioned application, it is more interesting to place emotions within a two or three-dimensional model of affective dimensions. Another reason for choosing the dimensional model of emotions, is that a dimensional model is very suitable for visualizing spontaneous, naturally occurring emotions because it allows for continuous description. In dimensional theory, emotions are usually placed within the dimensions valence (from positive to negative) and arousal (from high to low) [16, 20]. In [3], a third dimension called stance (from open to close) was added.

In Fig. 1, the 16 emotions that the intelligent agent can detect in speech are visualized according to the circumplex model of emotion [16, 20]. In this figure, the labels "anger", "anxiety", "irritation", "guilt", "happiness" and "content" correspond to the states "angry", "alarmed", "annoyed",

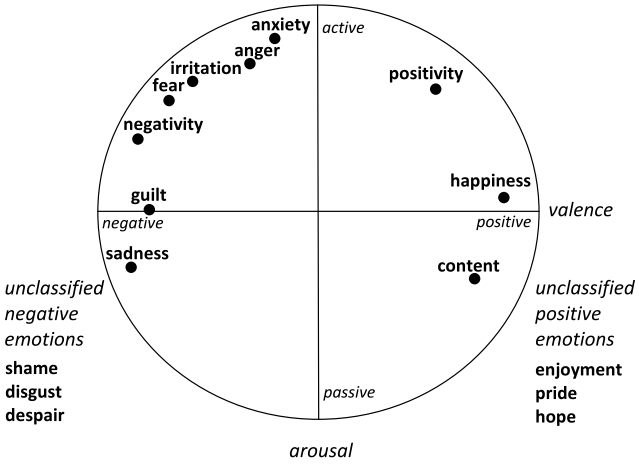

Fig. 1 Visualization of the 16 emotions according to the circumplex model of emotions in [16] and [20]

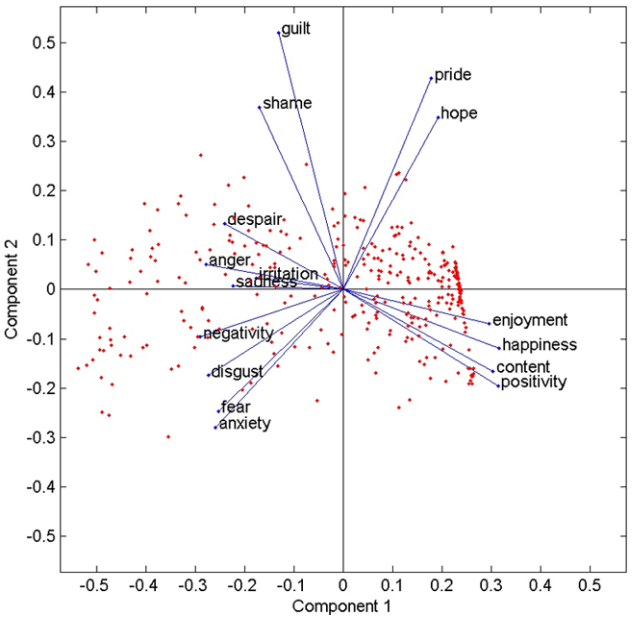

Fig. 2 First two principal components of all emotion vectors of original experiment

"guilty", "happy", and "content" as described in [16], respectively, while "fear", "negativity", "sadness" and "positivity" correspond to respective states described in [20]. The emotions outside the circle, shame, disgust, despair, enjoyment, pride and hope were not modeled in [16] or [20] and were left outside the circle. The valence of these emotions is known: shame, disgust and despair are negative emotions and enjoyment, pride and hope are positive emotions. The arousal is not clear beforehand according to the circumplex models. It will be interesting if the emotions in the current and original experiment correspond with the dimensional modeling of the circumplex model. If they differ, it can be seen how the ratings of people listening to a voice differ from the ratings of the speakers about their internal feelings.

5.2 PCA analysis of research data

PCA analysis has been performed on all emotion vectors of the 111 participants of the original experiment from which the emotion vectors and recordings of 75 participants (in the experimental conditions) were used in the current experiments. Each participant rated his/her emotions 3 times during the original experiment and all three emotion vectors for each participant were used in the PCA analysis. Also, PCA Analysis has been performed on the judged emotion vectors of the current Experiment 1. Before performing PCA a number of transformations were applied to the data. First, the range of scores from [1–10] was transformed to [0–9] just by subtracting 1 from each score. Second, every vector of 16 scores was divided by their sum, so that every vector would sum up to 1: $s_i = s_i/(s_1 + s_2 + \cdots + s_{16})$, for $i = 1, \ldots, 16$. The motivation for this normalization, is that it corrects for the response bias people have. Some people tend to score at the ends of the Likert scale, some people tend to score around the middle of the Likert Scale. To correct for this, everybody can now score within the same interval [0, 1]. Third, each emotion was standardized: calculate for each emotion e its mean value mean(e) and standard deviation std(e) and apply the formula: z score(s) $= (s - \text{mean}(e))/\text{std}(e)$, where s is a score of e given by a user.

The results of the first PCA analysis are shown in Fig. 2 and Table 14. Figure 2 shows that the first principal component corresponds with the valence dimension of the circumplex dimensional model of emotions; all negative emotions are on the left and all positive emotions are on the right. The second principal component seems to correspond with the energy dimension of the circumplex model for the emotions, upside down. It could make sense if fear, anxiety and disgust are the more active emotions and guilt and shame are more passive (more internal feelings). The same holds for the positive emotions: pride and hope could be the more passive emotions (more internal feelings) and enjoyment, happiness, content and positivity the more active emotions.

The second PCA analysis was performed in the same way for the ratings of the participants in Experiment 1. Figure 3 and Table 15 show the results. It can be seen that when one rates another person, different results appear than when the persons rate themselves. Also in Fig. 3, the first principal component corresponds to the valence dimension of the circumplex model: all negative emotions are on the left and all positive emotions are on the left. It can also be seen that it was hard to differentiate in the intensities between the negative emotions: all negative emotions, besides sadness and negativity, seem to lie close to each other. For positive emotions, there is a clearer division in emotions: content, happiness and positivity were rated different than hope, enjoyment and pride. Also, positive emotions were rated the op-

Table 14 Factor loadings of the first three principal components, belonging to Fig. 3

Emotions	Factor loading C_1	Factor loading C_2	Factor loading C_3
Negativity	−0.29	−0.10	0.09
Anger	−0.28	0.05	0.16
Disgust	−0.27	−0.17	0.01
Anxiety	−0.26	−0.28	−0.39
Fear	−0.25	−0.25	−0.38
Despair	−0.24	0.13	−0.15
Sadness	−0.22	0.01	0.03
Irritation	−0.18	0.02	0.57
Shame	−0.17	0.37	0.31
Guilt	−0.13	0.52	−0.03
Pride	0.18	0.43	−0.36
Hope	0.19	0.35	−0.23
Enjoyment	0.30	−0.07	−0.00
Content	0.30	−0.17	0.11
Positivity	0.31	−0.20	0.12
Happiness	0.32	−0.12	0.12
Explained variance	$C_1 = 0.49149$	$C_1 + C_2 = 0.58229$	$C_1 + C_2 + C_3 = 0.65053$

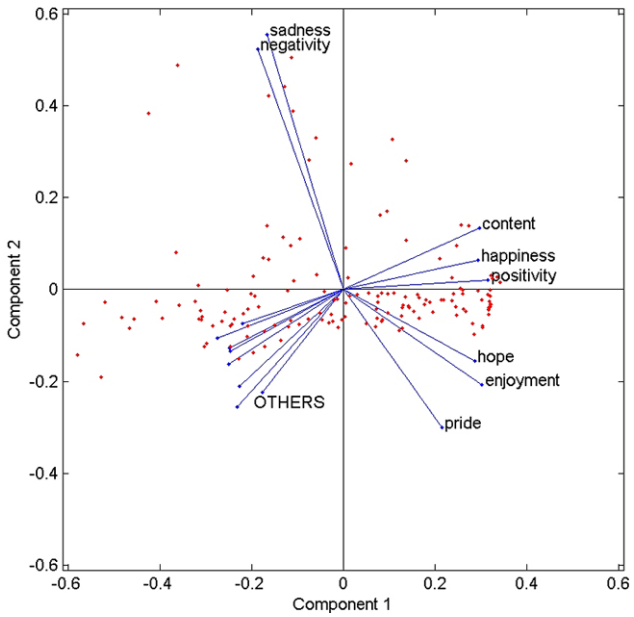

Fig. 3 First two principal components of all emotion vectors of speakers in Experiment 1 (subset of the emotion vectors in Fig. 2), but then scored by others: the 14 participants of Experiment 1

posite of the speakers own ratings. Again, it seems that the second principal component corresponds modestly with the energy dimension of the circumplex model, upside down. The most important conclusions are that the speakers' own ratings of their own feelings corresponds best to the circumplex model of emotions, which in a way validates the felt emotions of the speakers and this technique to visualize them in the envisioned application. Second, it has been also validated that another person's rating of your emotion can be very different than how you would rate your own feeling. Therefore, it is not a good idea to let your feelings be judged by other people, but to be judged by the intelligent agent that can learn your feelings, from your own feedback.

6 Conclusion

In this research, the goal was to develop a system that will automatically measure changes in the emotional state of a speaker, by analyzing his/her voice. Natural human speech was recorded in a laboratory study, and manually divided and labeled into meaningful pieces. In total, 207 speech features were extracted. The Random Forests algorithm was used to address a classification problem (out of 2 recordings, which one has the highest emotional value?) and a regression problem (given 2 recordings, estimate the change of each emotional state).

Results show that predicting the direction of change of emotions can be done about 7 % better than the baseline (the most frequent class label), while predicting the change of intensity, measured by the Mean Squared Error, can be done about 9.7 % better than the baseline (the mean value of change). Moreover, it turned out that changes of intensity in negative emotions are more predictable than changes in positive emotions: the relative error reduction rate for these two groups was 11.2 % and 7.1 %, respectively. At first sight, these error reductions could seem relatively small, but in fact they are not. Given the modest size of the training set (only 3 recordings per person) these improvements are quite

Table 15 Factor loadings of the first three principal components, belonging to Fig. 3

Emotions	Factor loading C_1	Factor loading C_2	Factor loading C_3
Despair	−0.27	−0.11	0.08
Anxiety	−0.25	−0.16	−0.42
Fear	−0.25	−0.13	−0.33
Disgust	−0.25	−0.13	−0.29
Anger	−0.23	−0.26	−0.14
Irritation	−0.23	−0.21	0.25
Shame	−0.22	−0.07	0.53
Negativity	−0.19	0.52	0.08
Guilt	−0.18	−0.22	0.42
Sadness	−0.17	0.56	−0.18
Pride	0.21	−0.30	−0.20
Hope	0.29	−0.16	−0.04
Happiness	0.29	0.06	−0.03
Content	0.30	0.13	0.04
Enjoyment	0.30	−0.21	−0.00
Positivity	0.32	0.02	0.04
Explained variance	$C_1 = 0.46868$	$C_1 + C_2 = 0.59101$	$C_1 + C_2 + C_3 = 0.6812$

remarkable, together with the fact that the performance is significantly better than a random guess. Moreover, in the field of affective sensing these improvements are quite substantial, for example see significant error reduction rates of only 7 % in [21].

In general, it is much easier to detect changes in negative emotions than in positive ones. This is beneficiary for the envisioned application, where an intelligent agent needs to detect negative mood, like anger, despair, fear and anxiety, to prevent the user from depression or to council the user into a positive mood. Occurrence of emotions like fear, shame, and despair can be predicted, on average, 29.2 % better than the baseline; change in their intensity can be predicted about 12.9 % better than the baseline. Some emotions are very difficult to predict. For example, guilt, happiness, hope, and pride seem to be not predictable at all (see Table 2).

The validation experiment investigated how human performance compared to the machine performance in 5 experiments. Experiments 1 and 2 investigated the following research questions: (1) Are humans better or worse than the machine algorithm in the classification of emotions? (out of 2 recordings, which one has the highest emotional value?), (2) Are humans better or worse than the machine algorithm in the regression problem? (Given 2 recordings, estimate the change of each emotional state), (3) Are humans better in detecting changing emotions in negative than positive emotions, than the machine algorithm?

The answer to the first research question is that the machine algorithm significantly outperformed the humans in Experiment 1 on 8 out of 16 emotions and in Experiment 2 on 12 out of 16 emotions. Happiness was the only emotion in both experiments where there was no significant difference found. This indicates, that overall, the machine algorithm is better in detecting changes in emotions than humans.

For question 2, results do not clearly indicate that the machine algorithm is better in predicting the size of change in intensity of emotions than humans. In Experiment 1, the machine algorithm did perform significantly better on 10 out of 16 emotions. For 5 more emotions a trend was found. In Experiment 2 though, the machine algorithm outperformed the humans significantly on only 3 emotions, from which the performance on sadness and anxiety correspond with that of Experiment 1, and a trend was found for 4 more emotions. Experiments 1 and 2 together, indicate that the machine algorithm is at least better in predicting the size of the change in anxiety and sadness, but for the other emotions, more experiments need to be conducted to make clear interpretations.

For question 3, no clear indications were found that the humans perform better in detecting changes in negative than positive emotions, like the machine algorithm. Experiments 3 to 5 investigated the following research questions: (4) Are humans better in predicting changing emotions in their own voice than somebody else's voice? (5) Are humans better in detecting changing emotions when they can also use semantic cues in the audio files, besides the acoustic cues? (6) Are humans better in detecting changing emotions when they learn how the speaker rated his own feelings than when they did not learn this? Results showed that over all emotions, the participants were always better in rating the emotional intensities of their own voice than that of others. These results indicate that it is easier to detect the intensities of emotions in one's own voice than that of a stranger.

Results of Experiment 4 do not confirm the expectation that humans can rate the emotional intensity better from recordings that include acoustic and semantic cues, than recordings with acoustic cues alone. The results show that the participants performed better in rating sadness, content, happiness, disgust, positivity, enjoyment, irritation, anger, pride and negativity in the non-"NAC" files conditions, but worse in rating fear, shame, guilt, despair, hope and anxiety. When looking at the averages over all emotions, the participants were marginally better in rating the emotional intensities of the non-"NAC" files than the "NAC" files.

Finally, the results of Experiment 5 demonstrate that over all emotions, humans can rate the emotional intensity of recordings when they have seen one of the speaker's own ratings with one of his recording marginally better than when not. In sum, all the results of the experiments validate that the machine algorithm performs better than humans, which is beneficiary for the envisioned application, where natural non-acted speech has to be processed by the intelligent agent.

In Sect. 5.2 all emotions were visualized by PCA analysis. The most important findings here were that the speakers' own ratings of their own feelings corresponds best to the circumplex model of emotions, which validates the felt emotions of the speakers and this technique to visualize them in the envisioned application. Second, it has also been validated that another person's rating of your emotion can be very different than how you would rate your own feeling. Therefore, it is not a good idea to let your feelings be judged by other people, but to be judged by the intelligent agent that can learn your feelings, by your own feedback.

7 Discussion

The biggest challenge in our research was a very scarce set of recordings: just 3 records per subject. We believe that with the increase of the size of available data, the accuracy of our models would dramatically improve. In practice, this should be easy to achieve: potential users of the final system will have to "tune" it to their specific voice and emotional states by providing numerous speech samples with labels, in the training phase.

Possible applications for the classification and regression algorithm were listed already in the introduction. Other applications are a warning system, a voice-based monitor of physiological functions and a computer game feedback monitor. For example, when a person is angry, aggressive or just furious, the smartphone could generate alerts or warnings like: "you are too excited to drive a car, operate heavy machinery or talk to your children" or "your blood pressure is probably too high right now". We hypothesize that these kind of warnings make the biggest impact if they are composed by the user itself. Concerning the voice-based monitor of physiological functions: we expect that there is a direct relation between voice characteristics and physiological states of a person, like breath rate, blood pressure, heart rate, sugar level and cholesterol level. We would like to experimentally verify and quantify this assumption, so we could build a very cheap monitoring device which would translate observed characteristic of speech into values of physiological parameters. Finally, the computer game monitoring system could use the learning algorithm to acquire information about the emotional state of the player via his/her voice to either verify if the intended effect of the current game level/environment is really there in the player, or to adjust the game level/environment to the current emotional state of the player.

The current research stems from the vision of an application, where speech is captured, while the person is communicating through a phone, to detect the current mood of a person. Most smart phones offer the possibility to capture the facial expression as well, via the video camera. If this could be incorporated in the envisioned application, multimodal emotion recognition through speech and facial expression would be possible. Processing videos requires a lot of computing power, though. Therefore, in the short term, real-time multimodal processing seems only feasible by applying facial recognition techniques to detect emotions in photos/still images taken of the user.

Moreover, in a follow up study, more audio files of the same speaker will be acquired for better accuracy of the system. Besides the speaker's self evaluation of his or her emotional state, it would be interesting to supplement the self-evaluation with a psychophysical measurement, like the Galvanic Skin Response. Finally, trying Support Vector Machines as an alternative modeling method is also part of future work.

References

1. Batliner A, Steidle S, Schuller B, Seppi D, Vogt T, Wagner J, Vidrascu L, Aharonson V, Kessous L, Amir N (2010) Whodunnit—searching for the most important speech feature types signalling emotion-related user states in speech. Comput Speech Lang. doi:10.1016/j.csl.2009.12.003
2. Breiman L (2001) Random forests. Mach Learn 45:5–32
3. Breazeal C, Brooks R (2005) Robot emotion: a functional perspective. In: Fellous J-M, Arbib MA (eds) Who needs emotions? Oxford University Press, New York
4. Castellano G, Kessous G, Caridakis G (2008) Emotion recognition through multiple modalities: face, body gesture, speech. In: Peter C, Beale R (eds) Affect and emotion in human-computer interaction. Lecture notes in computer science, vol 4868. Springer, Berlin, pp 92–103
5. Duda RO, Hart P, Stork D (2000) Pattern classification, 2nd edn. Wiley, New York

6. Ekman P (1992) An argument for basic emotions. Cogn Emot 6:169–200
7. Fredrickson BL, Mancuso R, Branigan C, Tugade M (2000) The undoing effect of positive emotions. Motiv Emot 24:237–258
8. Frijda NH (2007) The laws of emotion. Lawrence Erlbaum Associates Publishers, Hillsdale
9. GAQ (2002) Geneva appraisal questionnaire. See: http://www.affective-sciences.org/system/files/page/2636/GAQ_English.PDF
10. Hastie T, Tibshirani R, Friedman J (2008) The elements of statistical learning, 2nd edn. Springer, New York
11. Krematsu M, Amanuma S, Hakura J, Fujita H (2008) An extraction of emotion in human speech using cluster analysis and a regression tree. In: Fujita H, Sasaki J (eds) Proceedings of the 10th WSEAS international conference on applied computer science. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, pp 346–350
12. Laukka P, Neiberg D, Forsell M, Karlsson I, Elenius K (2011) Expression of affect in spontaneous speech: acoustic correlates and automatic detection of irritation and resignation. Comput Speech Lang 25:84–104
13. Li X, Tao J, Johnson M, Soltis J, Savage A, Leong K, Newman J (2007) Stress and emotion classification using jitter and shimmer features. In: IEEE international conference on acoustics, speech and signal processing (ICASSP 2007), pp 1081–1084
14. van der Maaten LJP, Postma E, van der Herik H (2009) Dimensionality reduction: a comparative review. Tilburg University technical report, TiCC-TR 2009-005
15. McIntyre G, Göcke R (2007) Towards affective sensing. In: Jacko JA (ed) Proc of the 12th international conference on human-computer interaction: intelligent multimodal interaction environments, part III (HCI'07). Lecture notes in computer science, vol 4552. Springer, Berlin, pp 411–420
16. Russel JA (1980) A circumplex model of affect. J Pers Soc Psychol 39:1161–1178
17. Schölkopf B, Smola AJ (2001) Learning with kernels. support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge
18. Tawari A, Trivedi M (2010) Speech based emotion classification framework for driver assistance system. In: Intelligent vehicles symposium (IV), 21–24 June 2010 IEEE Press, New York, pp 174–178. doi:10.1109/IVS.2010.5547956
19. Vogt T, André E, Wagner J (2007) Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. In: Jacko JA (ed) Proc of the 12th international conference on human-computer interaction: intelligent multimodal interaction environments, part III (HCI'07). Lecture notes in computer science, vol 4552. Springer, Berlin, pp 75–91
20. Yik M, Russel J, Steiger J (2011) A 12-point circumplex structure of core affect. Emotion 11(4):705–731
21. Zhang C, Wu J, Xiao X, Wang Z (2006) Pronunciation variation modeling for Mandarin with accent. In: Proceedings of ICSLP'06, Pittsburgh, USA, pp 709–712

C. Natalie van der Wal is an assistant professor in Artificial Intelligence at VU University Amsterdam. She works in the Agent Systems Research Group that investigates methods and techniques for modeling and analysis of agent systems in the area of human-oriented Ambient Intelligence. She finished her dissertation entitled: "Social Agents: Agent-Based Modeling of Intergrated Internal Social Dynamics of Cognitive and Affective Processes" in 2012. She is a multidisciplinary researcher that received her Masters degree (cum laude) in Artificial Intelligence in 2009, graduating on the topic of 'Modeling Agent-Based Support Systems for Group Emotion and Group Development'. She also received a Bachelors Degree in Cognitive and Clinical Neuropsychology in 2007 and a Masters Degree in Media and Culture in 2003. Her research interests are: modeling dynamics of agent systems in practical application areas, social diffusion of emotion and habits, neurological underpinnings of habits, promoting a healthy lifestyle with the help of technology. Her e-mail address is: c.n.vander.wal@vu.nl and her Web-page can be found at http://www.few.vu.nl/~cwl210/.

Wojtek Kowalczyk worked for more than 20 years as a part-time Assistant Professor in Artificial Intelligence at VU University Amsterdam, where he was responsible for developing numerous courses, including Neural Networks, Machine Learning, Data Mining Techniques and supervising more than 100 Master students in the fields of Artificial Intelligence and Business Analytics. Additionally, as a data mining consultant, he worked on numerous projects for banking, insurance, retail, transport, telecom and health care sectors, developing numerous self-learning systems for fraud detection, transaction monitoring, predicting traffic congestions, modeling customer behavior, tuning cochlear implants. Since 2012 he is with Leiden Institute of Advanced Computer Science, Leiden University, www.liacs.nl/~wojtek/.