Arousal and Valence prediction in spontaneous emotional speech: felt versus perceived emotion

Khiet P. Truong[1,2], *David A. van Leeuwen*[2], *Mark A. Neerincx*[2], *and Franciska M.G. de Jong*[1]

[1]University of Twente, Human Media Interaction, Enschede, The Netherlands
[2]TNO Defence, Security, and Safety, Soesterberg, The Netherlands
{k.p.truong,f.m.g.dejong}@ewi.utwente.nl, {david.vanleeuwen,mark.neerincx}@tno.nl

Abstract

In this paper, we describe emotion recognition experiments carried out for spontaneous affective speech with the aim to compare the added value of annotation of felt emotion versus annotation of perceived emotion. Using speech material available in the TNO-GAMING corpus (a corpus containing audiovisual recordings of people playing videogames), speech-based affect recognizers were developed that can predict Arousal and Valence scalar values. Two types of recognizers were developed in parallel: one trained with felt emotion annotations (generated by the gamers themselves) and one trained with perceived/observed emotion annotations (generated by a group of observers). The experiments showed that, in speech, with the methods and features currently used, observed emotions are easier to predict than felt emotions. The results suggest that recognition performance strongly depends on how and by whom the emotion annotations are carried out.

Index Terms: emotion, emotional speech database, emotion recognition

1. Introduction

In emotion recognition research, ground truth labels to be used for the development of emotion recognizers, are difficult to acquire and are to a certain extent subjective. There is, in general, no discussion about *who* is speaking or what *language* he or she is speaking, but people do not always agree on the speaker's *emotional state*. Hence, the labelling (annotation) of spontaneous expressive corpora remains a major topic in emotion research. One could assume that the closest approximation of 'ground truth' in emotion labelling is to ask the persons who have undergone the emotion to assign labels according to what they themselves *felt*. However, the majority of spontaneous emotion corpora contain emotion annotations that are generated by (naive) observers who can only label the *perceived* emotions. Only a small number of studies has investigated the use of annotations that are made by the subject who has undergone the emotion him/herself for expressive corpora. Auberge et al. [1] proposed to use 'auto-annotation', annotation performed by the subject him/herself, as an alternative method to label expressive corpora. The subjects were asked to label what they felt rather than what they expressed. There were no conclusive results: they concluded that 'felt'-annotations or 'expressed'-annotations both have their strengths and weaknesses. In Busso and Narayanan [2], the expression and perception of emotions were studied and 'self'-assessments of emotion were compared to assessments made by observers: the authors found a mismatch between felt and perceived emotions. The 'self'-raters appeared to assign their own emotions to more specific emotion categories which led to more extreme values in the Arousal-Valence space. In Truong et al. [3], we have also concluded that there are discrepancies between 'self' and perceived emotion assessments.

So far, we have not seen studies (to the best of our knowledge) that investigate whether these felt emotions, as labelled by the persons who have undergone these emotions themselves, can be predicted just as well as observed emotions. For some researchers, the ultimate goal is to develop a machine that can recognize one's felt emotions. From an emotion recognition perspective, it is important to know how the emotion signals were labelled and by whom. Hence, we developed, in parallel, two speech-based affect recognizers that can predict Arousal and Valence scalar values: one that is trained to detect felt emotions, and one that is trained to detect perceived emotions. The aim of this paper is to compare these speech-based affect recognizers' abilities to recognize felt or perceived emotion.

This paper is organized as follows. For the development of our recognizers, we used the TNO-GAMING corpus which is described in Section 2. We describe the method and speech features used to develop the recognizers in Section 3. The experimental setup is explained in Section 4, and the results of the experiments are presented in Section 5. Finally, in Section 6, we discuss the results and conclusions.

2. The TNO-GAMING corpus

For the development of the Arousal and Valence predictors, we used speech data from the TNO-GAMING corpus.

2.1. Audiovisual recordings

The TNO-GAMING corpus (see also [4, 3, 5]) contains audiovisual recordings of expressive behavior of subjects (17m/11f) playing a video game (*Unreal Tournament*). Speech recordings were made with high quality close-talk microphones. The audio of the game itself was played through headphones. Recordings of facial expressions were made with high quality webcams (Logitech Quickcam Sphere). In addition, the video stream of the game itself was also captured and stored at a rate of 1 frame per second. The participants played the video game twice in teams of 2 against 2. Expressive vocal and facial behavior of the participants was stimulated by 1) asking the participants to bring a friend as teammate, 2) granting bonusses to the team with the highest score and 'best' collaboration, and 3) generating suprising events during the game, e.g., sudden deaths, sudden appearances of monsters, and hampering mouse and keyboard controls.

2.2. Annotations performed by gamers themselves

One of the key characteristics of the TNO-GAMING corpus is that it is annotated by the gamers themselves. After each gaming session, the participants annoted their own emotions in two different ways by 1) choosing one of the twelve available emotion categories (Happiness, Boredom, Amusement, Surprise, Malicious Delight, Excitement, Fear, Anger, Relief, Frustration, Wonderment, and Disgust), and 2) giving Arousal and Valence ratings each 10 seconds (on scales ranging from -1 to 1). In our analyses, we only used the Arousal and Valence ratings. In this dimension-based continuous annotation procedure, the participants (who were offered the audiovisual recordings of the face and voice, and the video stream of the game) were asked to rate their own *felt* emotion on Arousal and Valence scales each 10 seconds; they could not pause or rewind the video. For this labelled data to be of use for the development of affect recognizers, we needed to post-process the data. Since the annotation was performed continuously by the participants in the dimension-based annotation, we needed to design a procedure that links the ratings given by the participants with certain spurts of speech (the ratings could have possibly been given at non-speech moments since the annotation was performed continuously). The post-processing procedure involved several steps: 1) detection and segmentation of the speech with a relatively simple energy-based silence detection algorithm (performed with Praat [6]), 2) manual word-level transcription of the speech (performed by the first author), and 3) synchronization of the speech segments obtained with the silence detection algorithm with the given Arousal and Valence ratings. This synchronization process was carried out as follows: for a maximum of N segments (we chose $N = 5$), check whether 1) the segment starts within a margin of T seconds (we chose $T = 3$) from the moment that the subject was requested to give the emotion judgement, and 2) the segment is labelled as non-silence by the silence detection algorithm. These procedures resulted in a total of 7473 rated speech segments, comprising a total duration of 186.2 minutes (mean of 1.5 s and standard deviation of 1.12 s) and a number of 1963 unique words. We refer to the annotations performed by the gamers themselves as SELF-annotations (and the gamers annotating their own emotions as SELF-raters).

2.3. Annotations performed by observers

A part of the corpus was also annotated by 6 (naive) observers who had not participated in the data collection procedure or the experiment described in [3]. From the total of 7473 speech segments, 2400 segments were selected (sampling the whole Arousal-Valence space of the SELF-annotations evenly) for re-annotation by the 6 naive observers (average age of 25.4 years). The 2400 segments have a total duration of 76 minutes (mean and standard deviation of 1.9 s and 1.2 s respectively).

The observers were asked to rate each audiovisual (pre-segmented) segment on the Arousal and Valence scale. Note that there are some differences with the SELF-annotation procedure: 1) the audiovisual segments are already segmented for the observers, 2) the observers can re-play the segment, and 3) the captured video stream of the game was not offered to the observers. To ensure that each segment was annotated by 3 different observers (in order to obtain more 'robust' emotion judgments), each observer annotated different overlapping parts of the data set. The data set of 2400 segments was divided into four parts, each part consisting of 624 segment. Each observer was assigned to two parts of this data set, and thus each observer annotated a total of 1248 segments. Of the 624 segments in each part, 24 segments occured twice and were used to assess the rating consistency of the observer (intra-rater reliability). For each observer, it took approximately 4 to 5 hours to complete the annotation of 1248 segments, including breaks. The annotations performed by these observers are referred to as OTHER.3-annotations, and the observers are referred to as OTHER.3-raters: '3' because each segment was rated by 3 different observers.

2.4. Speech material used in experiments

To recapitulate: 2400 segments were annotated by the gamers themselves and by observers, and hence, we can use two types of references: one that is based on SELF-ratings and one that is based on OTHER.3-ratings. The OTHER.3-ratings (a 3 by 2400 matrix) represent the 3 different (Arousal and Valence) ratings that each of the 2400 segment has (due to three different observers). In order to obtain a reference annotation of observers (1 Arousal and Valence rating per segment, a 1 by 2400 matrix), we averaged the 3 different ratings. These ratings are referred to as the OTHER.AVG-ratings and can be used, in parallel with the SELF-ratings, as reference for the development of affect recognizers.

The distribution of the two different types of references, SELF-ratings and OTHER.AVG-ratings are plotted in 2D-Histograms and shown in Fig. 1 and Fig. 2. These plots show that the observers judged the observed emotions much less extreme than the SELF-raters do: the OTHER.AVG-ratings are mostly located in the Neutral area. However, the pull towards Neutrality is also caused by the averaging process.

2.5. Analysis of felt and perceived emotion annotations

How consistent are the raters in their emotion annotations? Due to practical limitations, we were not able to assess the consistency of the SELF-raters, but for the 6 observers, we were able to assess their intra-rater consistencies. For the agreement computations, we used Krippendorff's α ([7]) and Pearson's ρ. For the computation of α (ordinal), all ratings were discretized into 5 classes (with boundaries at -0.6, -0.2, 0.2, and 0.6); we refer to this α as $\alpha_{\mathrm{ord},5}$). The observers obtained an averaged $\alpha_{\mathrm{ord},5}$ intra-reliability of 0.80 and 0.48, on a scale from -1 to 1, for Valence and Arousal respectively. It seems that the observers were more consistent in their Valence judgements than in their Arousal judgements. In Table 1, the agreement figures

Table 1: *Agreement between* SELF-*ratings ('felt') and* OTHER.AVG-*ratings ('perceived').*

	$\alpha_{\mathrm{ord},5}$	Pearson's ρ
Arousal	0.27	0.33
Valence	0.36	0.41

between the SELF-ratings ('felt') and OTHER.AVG-ratings ('perceived') are presented. These relatively low agreement figures indicate and confirm that there are discrepancies between felt and perceived emotion (see also [3, 2]), which are also visible in Fig. 1 and Fig. 2.

3. Method and Features

The method and features used to develop the speech-based affect recognizers are described here.

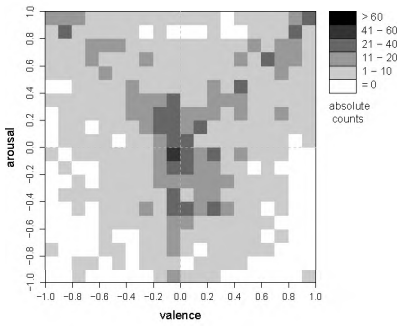

Figure 1: *2D Histogram: the distribution of the 2400 selected speech segments in the Arousal-Valence space, based on the* SELF*-ratings.*

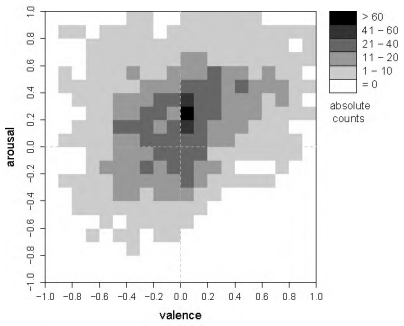

Figure 2: *2D Histogram: the distribution of the 2400 selected speech segments in the Arousal-Valence space, based on the* OTHER.AVG*-ratings.*

3.1. Support Vector Regression

Since our goal is to predict scalar values rather than discrete classes, we used a learning algorithm based on regression. We used Support Vector Regression (SVR, see [8]) to train regression models that can predict Arousal and Valence scalar values on a continuous scale. Similar to Support Vector Machines ([9]), SVR is a kernel-based method and allows the use of the kernel trick to transform the original feature space to a higher-dimensional feature space through a (non-linear) kernel function. We used ϵ-SVR available in *libsvm* ([10]) to train our models. In SVR, a margin ϵ is introduced and SVR tries to construct a discriminative hyperplane that has at most ϵ deviation from the original training samples. In our emotion prediction experiments, the RBF kernel function was used, and the parameters c (cost), ϵ (the ϵ of the loss function), and γ were tuned on a development set. The parameters were tuned via a simple grid search procedure that evaluates all possible combinations of c (with exponentially growing values between 2^{-4} and 2^{4}), ϵ (with exponentially growing values between 10^{-3} and 10^{0}), and γ (with exponentially growing values between 2^{-10} and 2^{2}).

3.2. Speech features

The acoustic feature extraction was performed with Praat ([6]). First, a voiced-unvoiced detection algorithm (available in Praat) was applied to find the voiced units. The features were extracted over each voiced unit of a segment. We made a selection of features based on previous studies (e.g., [12, 11]), and grouped these into features related to *pitch* information, *energy/intensity* information, and information about the *distribution of energy in the spectrum.* The spectral features MFCCs as commonly used in automatic speech recognition were also included. And finally, global information calculated over the whole segment (instead of per voiced unit) about the speech rate and the intensity and pitch contour was included. An overview of the features used is given in Table 2.

The majority of our acoustic features were measured per voiced unit. Subsequently, the features extracted on voiced-unit-level were aggregated to segment-level by taking the **mean**, **minimum**, and **maximum** of the features over the voiced units. Hence, we obtained per segment a feature vector with $(3 \times (4 + 4 + 5 + 24)) + 6 = 117$ dimensions. These features were normalized by transforming the features to z-scores:

Table 2: *Acoustic features used for emotion prediction with SVR.*

Features (with the number of features used in brackets)	
Pitch-related ($N = 4$)	mean, standard deviation, range (max-min), mean absolute pitch slope
Intensity-related ($N = 4$)	Root-Mean-Square (RMS), mean, range (max-min), standard deviation
Distribution-energy-in-spectrum-related ($N = 5$)	slope Long-Term Averaged Spectrum (LTAS), Hammarberg index, standard deviation, centre of gravity (cog), skewness
MFCCs ($N = 24$)	12 MFCC coefficients, 12 deltas (first order derivatives)
other ($N = 6$)	speech rate1, speech rate2, mean positive slope pitch, mean negative slope pitch, mean positive slope intensity, mean negative slope intensity

$z = (x - \mu)/\sigma$, with μ and σ calculated over a development set.

4. Experiments

Two speech-based affect recognizers were trained and tested in parallel: one that is trained to detect felt emotion and one that is trained to detect perceived emotion. In this Section, we describe the experimental setup and the evaluation metrics used.

4.1. Experimental setup

The automatic emotion prediction experiments (we use the term 'prediction' to emphasize the fact that we are predicting scalar values rather than discrete categories) were carried out speaker-independently, and separately for female and male speakers. We performed N-fold cross-validation, where in each fold, one specific speaker was held out for testing. The dataset of 2400 segments (1048f/1352m) was divided into training, development and test sets, where the training and test sets are disjoint. The splits in training/development/test are roughly 80%/10%/10% and 87%/8%/5% for female and male speakers respectively. The test set consists of speech segments from a specific speaker that is excluded from the training and development set. The

development set is comprised of randomly picked segments, drawn from the remaining segments after the test speaker has been filtered out.

The development set is used for parameter tuning and feature normalization (see Section 3). In parameter tuning, the parameter set that achieves the lowest error rate (e_{avg}, see Section 4.2), averaged over N folds, is selected to use in the final testing.

We performed two types of prediction experiments. One is based on the SELF-annotations, and the other one is based on the OTHER.AVG-annotations. With these two experiments, we investigate whether 'felt' or 'observed/perceived' emotions can be best predicted automatically.

4.2. Evaluation metrics

Because there are various evaluation metrics applicable to this emotion prediction task, we report several evaluation metrics. Firstly, we use a relatively simple evaluation metric that measures the absolute difference between the predicted output and the reference input (also used in [13]): $e_i = |x_i^{\text{pred}} - x_i^{\text{ref}}|$. We report the e_{avg} which is obtained by averaging over N segments: $e_{\text{avg}} = \frac{1}{N}\sum_i^N e_i$. The lower e_{avg}, the better the performance. Secondly, as human-machine agreement measure, we report Krippendorff's $\alpha_{ord,5}$ to allow comparison with human performance. Finally, Pearson's ρ is reported.

5. Results

The results of the Arousal and Valence prediction experiments are presented in Table 3. Some interesting observations can be made on the basis of these results. Firstly, we can observe that the performance obtained with the SELF-annotations as reference is much lower than when OTHER.AVG-annotations are used. This suggests that it is easier to predict perceived affect than felt affect.

Table 3: *Results of Arousal (=A) and Valence (=V) prediction experiments. The baseline results are obtained with a predictor that always predicts Neutrality.*

	Reference	Test SVR prediction			Baseline	
		e_{avg}	$\alpha_{\text{ord},5}$	ρ	e_{avg}	$\alpha_{\text{ord},5}$
A	SELF	0.41	0.22	0.25	0.45	-0.07
	OTHER.AVG	0.21	0.42	0.55	0.31	-0.18
V	SELF	0.36	0.10	0.18	0.36	-0.01
	OTHER.AVG	0.26	0.28	0.41	0.28	0.00

Secondly, Arousal can be much better predicted than Valence. Thirdly, although the predictors perform better than the baseline (a predictor that always predicts Neutrality), the relatively low agreement and correlation measures between the machine's predictions and the human judgements indictate that the performance in general seems to be rather moderate from a classification perspective.

6. Discussion and Conclusions

To summarize, the results of the experiments indicate that felt emotions are hard to predict using current recognition technology. It suggests that currently, we can only recognize expressed emotions that are perceivable by observers. The OTHER.AVG-annotations were obtained in a slightly different way than the SELF-annotations (due to practical limitations); these differences (see Section 2.3) may have resulted in slightly noisier SELF-annotations which possibly negatively affected prediction performance. Here we can remark that the SELF-annotations are by design all made by different annotators, and hence, we are doing "annotator-independent" prediction, whereas in the OTHER.AVG condition, the annotators are drawn from the same pool in training and testing. Furthermore, in general, the acoustic Arousal and Valence predictors appear to perform rather moderately from a classifier's perspective (although it should be noted that we did not optimize performance by e.g., feature selection). In future research, we will investigate more closely the relation between human and machine performance, and the relation between the quality of annotation and machine performance. In addition, the audiovisual recordings can be investigated for a multimodal analysis of affect, i.e., combining facial and vocal expressive behavior.

7. Acknowledgements

We would like to thank the 6 annotators who generated the perceived emotion annotations: Coraline, Frank, Piet, Ralph, and Thijs (trainees at TNO), and Ate. This work was supported by MultimediaN, a Dutch BSIK-project.

8. References

[1] Auberge, V. and Audibert, N. and Rilliard, A., "Auto-annotation: an alternative method to label expressive corpora", in Proceedings of LREC, 2006.

[2] Busso, C. and Narayanan, S. S., "The expression and perception of emotions: Comparing Assessments of Self versus Others", in Proceedings of Interspeech, 257–260, 2008.

[3] Truong, K.P. and Neerincx, M.A. and Leeuwen, D.A. van, "Assessing Agreement of Observer- and Self-Annotations in Spontaneous Multimodal Emotion Data", in Proceedings of Interspeech, 318–322, 2008.

[4] Merkx, P.P.A.B. and Truong, K.P. and Neerincx, M.A., "Inducing and measuring emotion through a multiplayer first-person shooter computer game", in Proceedings of Computer Games Workshop, 2007.

[5] Truong, K. P. and Raaijmakers, S., "Automatic Recognition of Spontaneous Emotions in Speech Using Acoustic and Lexical Features", in Proceedings of MLMI, 161–172, 2008.

[6] Boersma, P. and Weenink, D., "Praat: doing phonetics by computer", Online: http://www.praat.org.

[7] Krippendorff, K., "Reliability in Content Analysis", Human Communication Research, 30(3):411–433, 2004.

[8] Smola, A. J. and Schölkopf, B., "A tutorial on support vector regression", produced as part of the ESPRIT Working Group in Neural and Computational Learning II, Online:http://www.svms.org/regression/SmSc98.pdf, 1998.

[9] Vapnik, V.N., The nature of statistical learning theory, Springer-Verlag, New York, USA, 1995.

[10] Chang, C.-C. and Lin, C.-J., "LIBSVM: a library for Support Vector Machines", Online: http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.

[11] Banse, R. and Scherer, K. R., "Acoustic profiles in vocal emotion expression", Journal of Personality and Social Psychology, 70:614–636, 1996.

[12] Ververidis, D. and Kotropoulos, C. "Emotional speech recognition: Resources, features, and methods", Speech Communication, 48(9):1162-1181, 2006.

[13] Grimm, M. and Kroschel, K. and Narayanan, S., "Support vector regression for automatic recognition of spontaneous emotions in speech", in Proceedings of ICASSP, 1085–1088, 2007.