Proc. ICASSP 2007
Honolulu, HI

SUPPORT VECTOR REGRESSION FOR AUTOMATIC RECOGNITION OF SPONTANEOUS EMOTIONS IN SPEECH

Michael Grimm, Kristian Kroschel

Institut für Nachrichtentechnik (INT)
Universität Karlsruhe (TH)
Karlsruhe, Germany

Shrikanth Narayanan

Speech Analysis and Interpretation Lab (SAIL)
University of Southern California (USC)
Los Angeles CA, USA

ABSTRACT

We present novel methods for estimating spontaneously expressed emotions in speech. Three continuous-valued emotion primitives are used to describe emotions, namely *valence*, *activation*, and *dominance*. For the estimation of these primitives, Support Vector Machines (SVMs) are used in their application for regression (Support Vector Regression, SVR). Feature selection and parameter optimization are studied. The data was recorded from 47 speakers in a German talk-show on TV. The results were compared to a rule-based Fuzzy Logic classifier and a Fuzzy k-Nearest Neighbor classifier. SVR was found to give the best results and to be suited well for emotion estimation yielding small classification errors and high correlation between estimates and reference.

Index Terms— Speech analysis, Speech processing, User interface human factors, User modeling.

1. INTRODUCTION

This paper addresses an important question that has emerged in recent discussions on automatic emotion recognition: how to estimate emotions under the conditions of (1) non-acted, spontaneous speech and (2) non-categorical, quasi-continuous emotional content. Emotion recognition is important for human-machine interaction applications. For example, it is a key ingredient in the design of humanoid robots [1], where "emotional intelligence" is being increasingly added to artificial intelligence design [2]. The focus of this paper is on robust automatic emotion recognition from spontaneous conversational speech, and specifically, in finding an appropriate estimation method that goes beyond current multiple classification techniques.

Most research on automatic emotion recognition using the speech signal tries to distinguish a small number of emotion categories, such as negative and non-negative [3] or the set of anger, disgust, fear, joy, sadness, surprise and neutrality [4]. However, in general the expression of emotions in natural speech is not binary, i.e., angry or happy, but may take any arbitrary value in between. To describe this continuum we conceive emotions to be composed of three attributes which we will call *emotion primitives* in the following. As proposed by Kehrein [5] these primitives are *valence*, describing the negative vs. positive nature of an emotion, *activation*, describing the excitation on a scale from calm to excited, and *dominance*, describing the appearance of the person on a scale from submissive or weak to dominant or strong. Without loss of generality, they can be normalized to take values in the range of $[-1, +1]$ each. Estimating emotions on a continuous-valued scale provides an essential framework for recognizing dynamics in emotions, tracking intensities in the course of time, and adapting to individual moods or personalities.

In our previous work we analyzed a rule-based Fuzzy Logic classifier to estimate the emotion primitives from the speech signal [6], resulting in a mean error of 0.33 and an average correlation of 0.67 between the estimates and the reference annotated subjectively by listeners. Other approaches subdivide the emotion space into a few subclasses per primitive, and then perform discrete categorization [7, 8, 9]. To the best of our knowledge there is no other work on directly estimating the continuous values of these primitives, while in emotional speech synthesis regression methods applied to emotion primitives have been found to be very suitable [10]. In this study we chose Support Vector Machines and k-Nearest Neighbor classifiers, since they were found to be among the best classifiers for emotion categorization [11, 4], and used them in a modified version of Support Vector Regression (SVR) and Fuzzy k-Nearest Neighbor classifiers, respectively, for emotion estimation. In particular, we chose SVR because it is based on a solid theoretical framework to minimize the structural risk and not only the training error (empirical risk) [12]. It allows for complex, non-linear regression while providing results very fast at runtime.

The rest of the paper is organized as follows. Section 2 briefly introduces the data we used, and it also describes the emotion evaluation by human listeners. Section 3 describes the pre-processing steps of feature extraction and feature selection. Section 4 presents the different classifiers used for continuous-valued emotion primitive estimation. Section 5 describes the results and discusses the different estimator outcomes. Section 6 contains the conclusion and directions for

This work was supported by grants of the Collaborative Research Center (SFB) 588 of the Deutsche Forschungsgemeinschaft (DFG).
Contact: grimm@int.uni-karlsruhe.de

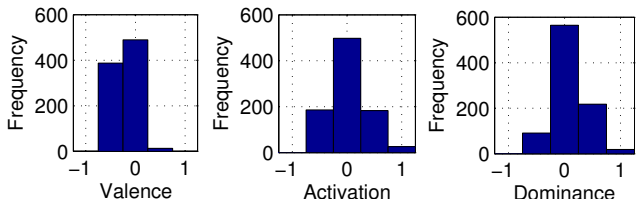

Fig. 1. Histogram of the emotions in the VAM corpus.

future work.
Note that the term "emotion" is a poorly (and broadly) defined term. It refers to a very complex inner state of a person including a wide range of mental and physical events. In the scope of this paper we understand the term "emotion" only as the visible part of this inner state that is transmitted through the speech signal and thus observable by a human receiver.

2. DATA

2.1. Data acquisition

For this study we used the *VAM corpus*, which contains data from a German TV talk-show in which several guests talk about personal issues such as friendship problems and fatherhood questions in a spontaneous, affective and unscripted manner. The dialogues were segmented into utterances. The signals were sampled at 16 kHz and 16 bit resolution.
Due to the topics, the data contains many negative emotions and few positive ones. In total the corpus contains 893 sentences from 47 speakers (11m/36f).
The emotion in each utterance was evaluated in a listener test (c.f. Sec. 2.2). Based on such a human evaluation, Fig. 1 shows the histogram of the emotions contained in the database. The attested emotion was taken as the reference for the automatic recognition, since assessment by the speakers themselves was not available.

2.2. Emotion evaluation

For evaluation we used an icon-based method based on *Self Assessment Manikins* [13] that yields one reference value $x_n^{(i)}$ for each primitive $i \in \{\textit{valence}, \textit{activation}, \textit{dominance}\}$. The listeners were played an utterance and they were asked to select the best describing image for each primitive afterwards. The individual listener ratings were averaged using confidence scores as described in [14]. One half of the database was evaluated by 17 listeners, the other by 6 listeners.
The average standard deviation in the evaluation was 0.29, 0.34, and 0.31 for *valence*, *activation*, and *dominance*, respectively. The mean correlation between the evaluators was 0.49, 0.72, and 0.61, respectively. Thus, *valence* was significantly more difficult to evaluate than *activation* or *dominance*. However, this result might also be an artefact of the correlation coefficient including the variance of the distribution, which is also smaller for *valence*.

3. PRE-PROCESSING

3.1. Feature extraction

In accordance with other studies on automatic emotion recognition we extracted prosodic features from the fundamental frequency (pitch) and the energy contours of the speech signals. The first and the second derivatives were also used. From these signals we calculated the following statistical parameters: mean value, maximum, minimum, range, median, 25% and 75% quartiles, and difference between the quartiles. In addition we used temporal characteristics such as speaking rate and pause to speech ratio, and spectral characteristics in 13 subbands derived from the Mel Frequency Cepstral Coefficients (MFCCs). In total 137 acoustic features were extracted. They were normalized to the range $[0, 1]$.

3.2. Feature selection

To reduce the large amount of acoustic features, we used the Sequential Forward Selection (SFS) technique for feature selection [15]. We found that, for each of the primitives and each of the classifiers, using 20 features was sufficient, and adding more features hardly improved the results. Compared to Principal Component Analysis, SFS gave slightly better results.

4. EMOTION PRIMITIVES ESTIMATION

In the following subsections we briefly describe the individual estimators. We refrain from calling them "classifiers" since the desired output is not a classification into one of a finite set of categories but an estimation of continuous-valued emotion parameters, the primitives $x_n^{(i)} \in [-1, +1]$, $i \in \{\textit{valence}, \textit{activation}, \textit{dominance}\}$.

4.1. Support Vector Regression

Support Vector Regression (SVR) is a regression method based on Support Vector Machines (SVM) [12, 16, 17]. While SVMs are mostly used to perform classification by determining the maximum margin separation hyperplane between two classes, SVR tries the inverse, i.e., to find the optimal regression hyperplane so that most training samples lie within an ϵ-margin around this hyperplane. Fig. 2 shows an example for SVR.
The mathematical formulation of the problem how to find an optimal regression hyperplane using a finite set of training samples can be found in [12, 17].
Non-linear regression is done in an efficient way by applying the *Kernel trick*, i.e., to replace the inner product in the solution by a non-linear kernel function. We used the following kernel functions: radial basis function (SVR-RBF) with $\sigma = 3.5$, polynomial kernel (SVR-Poly) with $d = 1$ and linear kernel (SVR-Lin).

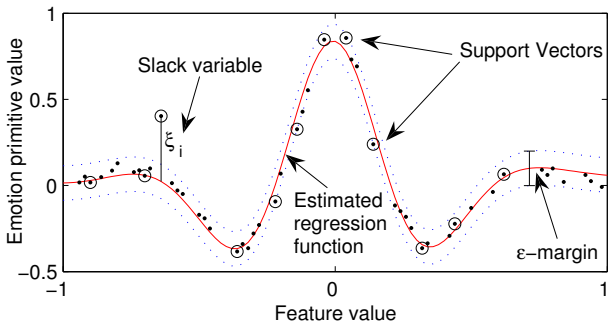

Fig. 2. Support Vector Regression example using the Kernel trick.

The design parameters of the SVR, C and ϵ, were chosen using a grid search on a logarithmic scale and a second, fine-grained search in the best region [18]. We chose $\epsilon = 0.2$ for all kernels and $C = 10$ and $C = 0.1$ for SVR-RBF and SVR-Poly or SVR-Lin, respectively. We used the *libsvm* implementation [19].

4.2. Fuzzy k-Nearest Neighbor estimator

The k-Nearest Neighbor (KNN) method determines the k closest neighbors of a query in the feature space and assigns the properties of these neighbors to the query [20]. In our case, the properties of these neighbors are the emotion primitive values, and these are averaged to get the final emotion estimate for a query. Due to this average, all neighbors have an influence on the estimate. This led us to calling the method *fuzzy* KNN.

The parameters to choose are p of the L_p distance and k. We chose $p = 2$ (Euclidean distance) and $k = 11$ since these parameters gave the best results for $p \in \{1, 1.5, \ldots, 4\}$ and $k \in \{1, 3, \ldots, 15\}$.

4.3. Rule-based Fuzzy Logic estimator

A rule-based Fuzzy Logic (FL) estimator has previously been used for automatic emotion primitive estimation [6]. The fuzzy logic captures well the nature of emotions, which in general is fuzzy in description and notation, while the rules in the inference system can be derived from expert knowledge or automatically by analysis of the relation between the acoustic features and the desired emotion estimates. We used a set of three linguistic, fuzzy variables for each primitive: *negative, neutral, positive* for *valence*; *calm, neutral, excited* for *activation*; and *weak, neutral, strong* for *dominance*.

The FL results can be seen as the baseline that has been achieved so far (c.f. Sec. 1). For each primitive, three membership functions were used, and each acoustic feature was related to each fuzzy variable by a rule that was automatically generated using the correlation between the acoustic feature and the emotion primitive [6]. Defuzzification into one crisp estimate was done using the centroid method.

	Valence	Activation	Dominance
SVR-RBF	0.13	0.15	0.14
SVR-Pol	0.14	0.16	0.15
SVR-Lin	0.13	0.16	0.15
FL	0.27	0.17	0.18
KNN	0.13	0.16	0.14

Table 1. Emotion primitives estimation results using different estimators: mean error.

	Valence	Activation	Dominance
SVR-RBF	0.46	0.82	0.79
SVR-Pol	0.39	0.80	0.77
SVR-Lin	0.37	0.80	0.77
FL	0.28	0.75	0.72
KNN	0.46	0.80	0.78

Table 2. Emotion primitives estimation results using different estimators: correlation coefficient.

5. RESULTS

All results were achieved using 10-fold cross-validation. Table 1 summarizes the mean linear error for each estimator, for each emotion primitive separately. Additionally, we report the correlation between the estimates of the emotion primitives and the references determined by the listener test (c.f. Sec. 2.2) in Table 2. The correlation shows the accuracy in the tendency of the estimates.

It can be seen that all primitives can be estimated with a small error in the range of 0.13 to 0.18, with the exception of *valence* when estimated using the FL estimator (0.27). However, the correlation between the estimates and the reference was significantly different for the individual emotion primitives. The correlation for *valence* was between 0.28 and 0.46 for the different estimators, and it was between 0.72 and 0.82 for *activation* and *dominance*. Thus the results imply very good recognition results for *activation* and *dominance*, and moderate recognition results for *valence*.

The best results were achieved using the SVR-RBF estimator with errors of 0.13, 0.15, and 0.14, and correlation coefficients of 0.46, 0.82, and 0.79 for *valence*, *activation*, and *dominance*, respectively. The Fuzzy KNN estimator performed almost as well as the SVR-RBF. The SVR-Poly and SVR-Lin estimators gave worse results for *valence* in comparison to almost as good results for *activation* and *dominance*. Similarly, the FL estimator gave even worse results for *valence* but still very good results for *activation* and *dominance*.

Fig. 3 shows the estimation results using the SVR-RBF estimator, ordered in ascending order of the values of the primitives. It reveals some information about the nature of the errors: While most of the estimates are located within a small margin around the references, a small number of very high or very low primitive values was occasionally underestimated.

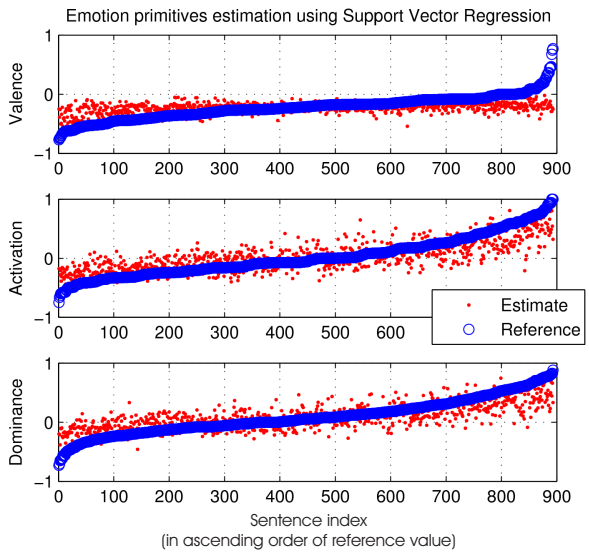

Fig. 3. Emotion primitives estimation results using SVR: direct comparison of estimates and reference.

6. CONCLUSION

We analyzed continuous-valued estimation of three emotion primitives, namely *valence*, *activation*, and *dominance*, using 20 acoustic features that were extracted from the prosody and the spectrum of spontaneous speech signals. For estimation, Support Vector Regression, Fuzzy Logic, and Fuzzy k-Nearest Neighbor methods were used. We found that the emotion primitives could be estimated with a small error of 0.13 to 0.15, where the range of values was [-1,+1], and a moderate (0.46, *valence*) to high (0.82/0.79, *activation/dominance*) correlation between estimates and reference. The error for *valence* estimation could be reduced by 52% compared to the Fuzzy-Logic baseline, and the error for *activation* and *dominance* was reduced by 12% and 22%, respectively.
Support Vector Regression gave the best estimation results. Note that while this algorithm is computationally much more demanding for initialization (finding the regression hyperplane), the KNN method requires more computational power at the actual estimation step due to the distance matrix that has to be calculated. The rule-based FL algorithm is computationally less demanding but gives clearly inferior results, at least for *valence*.

Future work will investigate designing a real-time system using the algorithms that were reported here. The advantage of continuous-valued estimates of the emotional state of a person could be used to build an adaptive emotion tracking system that is capable to adapt to individual personalities and long-term moods.

7. REFERENCES

[1] "Humanoid Robots – Learning and Cooperating Multimodal Robots," Collaborative Research Center of the Deutsche Forschungsgemeinschaft, http://www.sfb588.uni-karlsruhe.de/, 2001.

[2] R.W. Picard, *Affective Computing*, MIT Press, Cambridge, MA, USA, 1997.

[3] C.M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *Proc. IEEE Automatic Speech Recognition and Understanding Wsh. (ASRU)*, 2001.

[4] B. Schuller, S. Reiter, et al., "Speaker Independent Speech Emotion Recognition by Ensemble Classification," in *Proc. Int. Conf. on Multimedia and Expo*, 2005, pp. 864–867.

[5] R. Kehrein, "The prosody of authentic emotions," in *Proc. Speech Prosody Conf.*, 2002, pp. 423–426.

[6] M. Grimm and K. Kroschel, "Rule-based emotion classification using acoustic features," in *Proc. Int. Conf. on Telemedicine and Multimedia Communication*, 2005.

[7] C. Yu, P.M. Aoki, and A. Woodruff, "Detecting user engagement in everyday conversations," in *Proc. Int. Conf. Spoken Lang. Processing*, 2004, vol. 2, pp. 1329–1332.

[8] L. Vidrascu and L. Devillers, "Real-life emotion representation and detection in call centers data.," in *Proc. Int. Conf. on Affective Computing and Intelligent Interaction*, 2005, pp. 739–746.

[9] N.F. Fragopanagos and J.G. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

[10] M. Schröder, "Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis," 2003, PhD Thesis, Universität des Saarlandes, Germany.

[11] S. Yacoub, S. Simske, et al., "Recognition of Emotions in Interactive Voice Response Systems," Tech. Rep., HP Laboratories Palo Alto, July 2003.

[12] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

[13] L. Fischer, D. Brauns, and F. Belschak, *Zur Messung von Emotionen in der angewandten Forschung*, Pabst Science Publishers, Lengerich, 2002.

[14] M. Grimm and K. Kroschel, "Evaluation of Natural Emotions Using Self Assessment Manikins," in *Proc. ASRU*, 2005, pp. 381–385.

[15] J. Kittler, "Feature set search algorithms," *Pattern Recognition and Signal Processing*, pp. 41–60, 1978.

[16] C. Campbell, *An Introduction to Kernel Methods*, pp. 155–192, Physica-Verlag Heidelberg, 2001, in: R.J. Howlett and L.C. Jain, "Radial Basis Function Networks 1".

[17] B. Schölkopf and A.J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, The MIT Press, Cambridge, MA, 2001.

[18] R. Chudoba, "Klassifikation von Emotionen mit Support Vector Machines," Diploma Thesis, Universität Karlsruhe (TH), Germany, 2006.

[19] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, http://www.csie.ntu.edu.tw/˜cjlin/libsvm.

[20] K. Kroschel, *Statistische Informationstheorie*, Springer Verlag Berlin, 4. edition, 2004.