interspeech 2008 incorporating SST 08

Abandoning Emotion Classes - Towards Continuous Emotion Recognition with Modelling of Long-Range Dependencies

Martin Wöllmer[1], *Florian Eyben*[1], *Stephan Reiter*[2], *Björn Schuller*[1], *Cate Cox*[3], *Ellen Douglas-Cowie*[3], *Roddy Cowie*[3]

[1]Technische Universität München, Institute for Human-Machine Communication, 80290 München, Germany
[2]EB Automotive GmbH, 91058 Erlangen, Germany
[3]Queen's University, School of Psychology, Belfast, BT7 1NN, UK
schuller@tum.de, woellmer@tum.de, eyben@tum.de, r.cowie@qub.ac.uk

10.21437/Interspeech.2008-192

Abstract

Class based emotion recognition from speech, as performed in most works up to now, entails many restrictions for practical applications. Human emotion is a continuum and an automatic emotion recognition system must be able to recognise it as such. We present a novel approach for continuous emotion recognition based on Long Short-Term Memory Recurrent Neural Networks which include modelling of long-range dependencies between observations and thus outperform techniques like Support-Vector Regression. Transferring the innovative concept of additionally modelling emotional history to the classification of discrete levels for the emotional dimensions "valence" and "activation" we also apply Conditional Random Fields which prevail over the commonly used Support-Vector Machines. Experiments conducted on data that was recorded while humans interacted with a Sensitive Artificial Listener prove that for activation the derived classifiers perform as well as human annotators.

Index Terms: Emotion Recognition, Sensitive Artificial Listener, LSTM

1. Introduction

Automatic emotion recognition from speech has in the past focused on identifying discrete classes of emotion, e. g. [1]. However, common sense and psychological studies suggest that the full spectrum of human emotion cannot be expressed by a few discrete classes. Emotion is better represented by continuous values on multiple attribute axes such as valence, activation or dominance [6]. A specific emotion thereby is represented by a point in a multi-dimensional coordinate space.

Research dealing with recognition of emotion as a continuum requires databases where emotion is continuously labeled regarding multiple attributes. Such databases [3] have been recently recorded by the HUMAINE project[1]. In the Sensitive Artificial Listener (SAL) database which is used in this work, not only emotional dimensions but also the dimension "time" is a quasi-continuum, since annotations for valence and activation are sampled every 10 ms. In this paper we introduce a new technique for continuous emotion recognition in a 3D space spanned by activation, valence, and time using Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) [7]. Recently LSTM-RNN have been successfully applied to speech recognition [5] and meeting segmentation [9]. In contrast to state-of-the-art approaches such as Support-Vector Regression (SVR), LSTM-RNN also model long-range dependencies between successive observations and therefore are suited to capture emotional history for adequate prediction of emotion in a three-dimensional space. The principle of modelling the temporal evolution of emotion is also evaluated for discrete labels. For this purpose the continuous labels for activation and valence are quantised and Conditional Random Fields (CRF) [8], which similarly to LSTM-RNN drop the independence assumption between successive class labels, are applied.

The next section introduces the database of induced emotional speech used in our experiments. Section 3 describes the set of acoustic features. In Section 4 the classification methods are explained. Finally, in Section 5 the results are presented.

2. Database

As database we use the induced Belfast Sensitive Artificial Listener data which is part of the final HUMAINE database [3]. We use a subset which contains 25 recordings in total from 4 speakers (2 male, 2 female) with an average length of 20 minutes per speaker. The data contains audio-visual recordings from natural human-computer conversations that were recorded through a SAL interface designed to let users work through a range of emotional states. [3] describes the database in more detail. Data has been labelled continuously in real time by 4 annotators with respect to valence and activation using a system based on FEELtrace [2]. The annotators used a sliding controller to annotate both emotional dimensions separately whereas the adjusted values for valence and activation were sampled every 10 ms to obtain a temporal quasi-continuum. To compensate linear offsets that are present among the annotators, the annotations were normalised to zero mean globally. Further, to ensure common scaling among all annotators, each annotator's labels were scaled so that 98% of all values are in the range from -1 to +1. Finally, the mean of all 4 annotators was computed, which is used as ground truth label in the experiments reported on in Section 5. The average Mean Squared Error (MSE) of the four human annotators with respect to the mean value is 0.08 for activation and 0.07 for valence. The 25 recordings have been split into turns using an energy based Voice Activity Detection. A total of 1,692 turns is accordingly contained in the database. The turns were once randomly divided into training (1,102 turns) and test (590 turns) splits for the experiments. Both sets contain all speakers, thus

[1]http://www.emotion-research.net/

Accepted after peer review of full paper

results are not speaker independent, which in turn would not be feasible with only 4 speakers. Labels for each turn are computed by averaging the frame level valence and activation labels over the complete turn. Apart from the necessity to deal with continuous values for time and emotion, the great challenge of emotion recognition on the SAL database is the fact that the system must deal with all data - as recorded - and not only manually pre-selected "emotional prototypes" as in practically any other database.

3. Feature Extraction

As acoustic features for emotion recognition, functionals of acoustic Low-Level Descriptors (LLD) are state of the art as proven by various current works, e. g. [1, 10]. These functionals are statistical properties derived from the LLD contours of the whole utterance to be classified. Thus, utterances of variable length can be mapped onto a feature vector of constant dimension. In this work the features as introduced in [11] are used. Table 1 shows the LLD to which the functionals maximum, position of maximum, minimum, position of minimum, mean, median, and standard deviation are applied. Thereby we used the principle of hierarchical functionals to compensate statistical outliers in long turns: the typical turn-wise functionals are supplemented by hierarchical functionals ("functionals of functionals", e. g. "mean of maxima") basing on a fixed length segmentation of 1 second. In [11] this novel strategy has proven to enable enhanced performance for emotion recognition.

Table 1: *Acoustic LLD used for computation of hierarchical functionals.*

Type	**LLD**
Time Signal	Elongation, Centroid, Zero-Crossing Rate
Energy	Log-Frame-Energy
Spectral	0-250 Hz, 0-650 Hz, Flux
	Roll-Off + δ, Centroid + δ
Pitch	F_0 (fundamental frequency)
Formants	F1-F7 Frequency + δ,
	+ Bandwidth + δ
Cepstral	MFCC 1-15 + δ + $\delta\delta$
Voice Quality	Harmonics to Noise Ratio (HNR)

In total 4,843 features have been extracted for each utterance. To investigate the effects of feature normalisation, six variations are evaluated: mean and variance standardisation (MVS), numeric normalisation to range -1 to +1 (NRM), and the combination of MVS and NRM (M+N) each applied to a) the complete data set (ALL) and b) the data of each speaker individually (SPK). In the ongoing the resulting features sets are named accordingly : MVS_{all}, NRM_{all}, $M{+}N_{all}$, MVS_{spk}, NRM_{spk}, and $M{+}N_{spk}$. For computation of the standardisation/normalisation coefficients only the training split was used.

To find features highly relevant to the task at hand a correlation-based feature search (CFS) basing on a Sequential-Forward-Floating-Search (SFFS) is performed for each target label and feature set individually using the corresponding training split.

4. Classification

4.1. Long Short-Term Memory Recurrent Neural Net

A major drawback of recurrent neural networks trained by back-propagation through time and other established methods is that they are not able to store information over a longer time period. Bridging such longer lags is difficult since error signals are likely to either blow up or vanish. With so-called Long Short-Term Memory (LSTM) cells as introduced by Hochreiter and Schmidhuber [7] it is possible to overcome the problem that events lying back in time tend to be forgotten. Instead of the hidden cells of a conventional recurrent neural net the LSTM-RNN consists of memory blocks which contain one or more memory cells (see Figure 1). In the middle of each cell there is a simple linear unit with a single self-recurrent connection whose weight is set to 1.0. Thus the current state of a cell is preserved throughout one time step. The output of one cell is

$$y^{c_j}(t) = y^{out_j}(t) h(s_{c_j}(t)) \quad (1)$$

whereas the internal state $s_{c_j}(t)$ can be calculated as

$$s_{c_j}(t) = s_{c_j}(t-1) + y^{in_j}(t) g(net_{c_j}(t)) \quad (2)$$

with the initial state $s_{c_j}(0)$ being 0. Due to this architecture salient events can be remembered over arbitrarily long periods of time. Cells can be combined to blocks sharing the input and output gate. In general, the LSTM-RNN architecture consists of three layers: an input, a hidden, and an output layer. The number of input layer nodes corresponds to the dimension of the feature vector. As hidden layer we used 8 blocks with four LSTM cells each. For the output layer one node is used, corresponding to either valence or arousal.

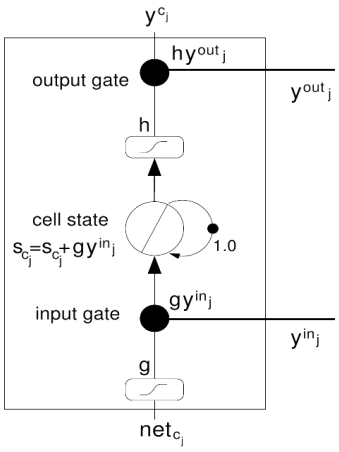

Figure 1: *Memory cell of an LSTM-RNN; s: states; y: data (inputs/outputs); g,h: transfer functions (sigmoid)*

4.2. Conditional Random Fields

As alternative to continuous class labelling with an LSTM-RNN, also discrete labelling using Conditional Random Fields (CRF) [8] was conducted in Section 5. Unlike generative models like the Hidden Markov Model, Conditional Random Fields do not assume that the observations are conditionally independent. This is advantageous whenever there are long-range dependencies between the observations. For CRF classification the continuous valence and arousal labels were quantised to four and seven levels each.

5. Results and Discussion

In this section recognition results for classification of valence and activation with four different classifiers are presented.

Apart from CRF we also use Support-Vector Machines (SVM) as a state-of-the-art method for discrete classification. Then we discuss the continuous predictors LSTM-RNN and SVR.

Table 2: *Recognition rates RR and RR^c (tolerating confusion between directly neighboring labels) for SVM classification of activation (A) and valence (V) with 4 (top part) and 7 (bottom part) discrete classes.*

Feature set	RR_A [%]	RR_A^c [%]	RR_V [%]	RR_V^c [%]
MVS_{all}	44.4	91.4	43.7	**87.3**
MVS_{spk}	41.9	89.2	32.7	79.2
NRM_{all}	**46.3**	**94.2**	44.4	86.9
NRM_{spk}	44.1	92.2	41.2	86.1
$M{+}N_{all}$	45.1	93.9	**44.6**	86.8
$M{+}N_{spk}$	45.5	92.5	41.2	86.6
MVS_{all}	26.4	67.6	22.4	55.4
MVS_{spk}	25.4	70.3	17.8	49.7
NRM_{all}	**30.8**	**73.4**	**24.6**	61.2
NRM_{spk}	28.1	71.4	22.4	**63.1**
$M{+}N_{all}$	30.2	73.2	24.4	62.0
$M{+}N_{spk}$	30.3	73.1	20.7	57.1

5.1. Support-Vector Machines

Table 2 shows results for SVM classification using the six different feature normalisation variants with 4 and 7 quantisation steps (discrete classes) for the continuous values of valence and activation. Depending on the accuracy requirements of the application confusions between neighboring classes may be tolerable. Thus, in addition to the standard recognition rate (RR) a second recognition rate (RR^c) is introduced where confusions between directly adjacent neighbouring classes are scored as correctly classified.

We would like to note at this point that the RR^c value for 4 classes has to be interpreted carefully, since many confusions are scored as correct, the observed high recognition rates can be expected. Thus, the same experiment was conducted with 7 classes. Then, each class represents a numerical range of width $\frac{1}{7} \approx 0.29$. Two classes represent a numerical range of width ≈ 0.58, being almost equivalent to using four classes. The RR^c rate for 7 classes is roughly around 70%, which is good considering the subjective nature of emotion. Results for activation are remarkably better than for valence which confirms the findings in [4] proving that even for humans valence is harder to identify than activation whenever linguistic information is not included.

5.2. Conditional Random Fields

Table 3 shows the corresponding results for CRF classification with the same number of quantisation steps and types of feature normalisation as in Section 5. As expected, due to additional modelling of temporal dependencies between the observations, CRF outperform SVM considering the best results for each case (printed in bold face in Table 2 and Table 3 respectively). The best recognition rates RR, both for valence and activation, are obtained with normalisations NRM_{all} and $M{+}N_{all}$ for both classifiers.

5.3. Support-Vector Regression

Continuous classifiers (predictors) require a different evaluation method. Instead of a recognition rate (percentage of correctly classified instances) the Mean Squared Error (MSE) between the prediction and the actual value is propagated herein. Larger deviations between actual and predicted have greater influence on the MSE than small errors. Therefore, the MSE in contrast to the Mean Linear Error (MLE) also enhances the accuracy of the result to some extent. The left part of Table 4 shows the MSE for valence and activation achieved with SVR classification.

Table 3: *Recognition rates RR and RR^c (tolerating confusion between directly neighboring labels) for CRF classification of activation (A) and valence (V) with 4 (top part) and 7 (bottom part) discrete classes.*

Feature set	RR_A [%]	RR_A^c [%]	RR_V [%]	RR_V^c [%]
MVS_{all}	46.4	91.5	43.7	**88.6**
MVS_{spk}	39.2	88.3	32.4	76.8
NRM_{all}	46.6	93.9	**45.6**	88.3
NRM_{spk}	44.1	**95.4**	36.6	79.0
$M{+}N_{all}$	**50.8**	95.0	44.9	88.0
$M{+}N_{spk}$	45.3	94.1	34.7	81.9
MVS_{all}	27.8	71.2	20.9	52.0
MVS_{spk}	28.0	69.2	18.3	45.8
NRM_{all}	**32.5**	77.5	**29.7**	**65.1**
NRM_{spk}	31.7	**77.8**	25.8	64.9
$M{+}N_{all}$	31.5	73.4	27.0	62.4
$M{+}N_{spk}$	26.1	70.9	21.7	62.7

Table 4: *MSE for SVR (left part) and LSTM-RNN (right part) prediction of activation (A) and valence (V) on 4 feature sets with respect to the mean label value of the 4 annotators.*

Feature set	SVR MSE_A	SVR MSE_V	LSTM MSE_A	LSTM MSE_V
MVN_{all}	0.12	0.19	0.10	0.22
MVN_{spk}	**0.10**	0.19	0.13	0.21
NRM_{all}	0.12	0.20	0.14	**0.18**
NRM_{spk}	0.12	**0.18**	0.12	0.25
$M{+}N_{all}$	0.12	0.20	0.14	0.20
$M{+}N_{spk}$	0.12	**0.18**	**0.08**	0.25

To visualise the correctness of the classified values and to compare the deviation of the predicted emotional dimensions with the degree of agreement between the 4 annotators, Figure 2 shows the classified value obtained with SVR (dashed broad grey line) as well as the minimum and the maximum of the annotated values (thin grey line) for all 590 test turns. The speaker boundaries (between speakers f1, f2, m1, m2) are marked with vertical dotted lines.

5.4. LSTM-RNN

The MSE for LSTM-RNN classification is shown in the right part of Table 4. While for valence the best performance obtained with LSTM-RNN is equal to the best performance with SVR (0.18 MSE), LSTM-RNN outperform SVR classification for activation. With feature normalisation type $M{+}N_{spk}$ the result for activation (MSE 0.08) is as good as human annotation (see Section 2), which again confirms that modelling long-range dependencies is advantageous for continuous emotion recognition. Also the broad black line in Figure 2, which shows the

classified values when using LSTM-RNN, illustrates the almost perfect result for activation: the classified activation value primarily lies within corridor defined by the deviation between the annotators (thin grey lines). As in the discrete case, valence is classified far less accurately because of the lack of linguistic information.

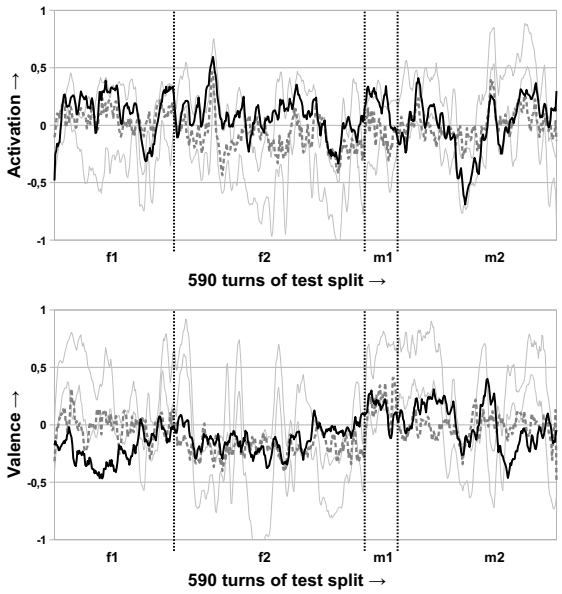

Figure 2: *LSTM-RNN (solid black line) / SVR (dashed grey line) classification of activation and valence (y-axis) over 590 test turns (x-axis); upper/lower grey line: maximum/minimum value of the 4 annotators.*

6. Conclusion and Outlook

In this work we introduced a novel strategy of emotion recognition operating in continuous three-dimensional space. In contrast to common static classification techniques which aim to distinguish discrete emotion classes and ignore temporal dependencies occurring in the evolution of affect, we operate in a quasi-continuous time domain with emotional dimensions "valence" and "arousal" continuously emerging over time. To capture long-range dependencies between the acoustic observations derived from hierarchical functionals of prosodic, spectral, and voice quality features, we model emotional history using Long Short-Term Memory Recurrent Networks which extend the principle of recurrent neural networks by including memory cells. The benefits of continuous LSTM-RNN modelling become evident in our experiments when comparing the strategy to state-of-the-art classification techniques like Support-Vector Regression: LSTM-RNN prevail over SVR and achieve a prediction quality which for "activation" is equal to human performance. To further prove the convenience of including long-range time dependencies between observations in quasi-time-continuous emotion recognition, we also evaluated the concept for discrete class labels obtained through quantisation of the continuous labels for valence and arousal. Again including emotional history, which in this case is done by applying Conditional Random Fields, prevails over standard techniques like Support-Vector Machines.

The classification performance for valence still is relatively low which derives from the fact that detecting valence from acoustic features alone is known to be a hard task, even for humans. To overcome this problem the integration of linguistic features which requires additional speech recognition is a possible future approach for robust continuous emotion recognition with LSTM-RNN. Further issues making the recognition of valence challenging can be found when considering the SAL database which was used in the experiments: only 4 speakers and 4 annotators with low inter-labeler-agreement are used, which is a low number compared to other databases for emotion recognition [6]. Apart from this, the SAL database consists not only of emotional prototypes with strong emotions. For future works it is interesting to apply the promising concept of modelling long-range dependencies in the temporal evolution of emotion on larger databases which are also labeled as a quasi-continuous 3D emotion space.

7. Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 211486 (SEMAINE).

8. References

[1] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to find trouble in communication. *Speech Communication 40*, pages 117–143, 2003.

[2] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. Feeltrace: an instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 19–24, 2000.

[3] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, , and K. Karpouzis. *The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data*, volume 4738, pages 488–500. Springer-Verlag Berlin Heidelberg, Lisbon, Portugal, 2007.

[4] E. Douglas-Cowie, L. Devillers, J. C. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox. Multimodal databases of everyday emotion: Facing up to complexity. In *Interspeech 2005*, pages 813–816, Lisbon, Portugal, 2005.

[5] A. Graves, D. Eck, N. Beringer, and J. Schmidhuber". Biologically plausible speech recognition with lstm neural nets". In *Proc. Bio-ADIT*, 2004.

[6] M. Grimm, K. Kroschel, H. Harris, C. Nass, B. Schuller, G. Rigoll, and T. Moosmayr. On the necessity and feasibility of detecting a driver's emotional state while driving. In *Proc. ACII 2007*, pages 126–238, Lisbon, 2007. IEEE.

[7] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, pages 1735–1780, 1997.

[8] J. Lafferty, A. McCallum, and F. Pereiar. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning 2001*, pages 282–289, 2001.

[9] S. Reiter, B. Schuller, and G. Rigoll. Hidden conditional random fields for meeting segmentation. In *Proc. ICME 2007*, pages 639–642, Beijing, China, July 2007.

[10] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. In *Proc. INTERSPEECH 2007, ISCA*, pages 2253–2256, Antwerp, Belgium, August 2007.

[11] B. Schuller, M. Wimmer, L. Mösenlechner, C. Kern, D. Arsic, and G. Rigoll. Brute-forcing hierarchical functionals for paralinguistics: a waste of feature space? In *Proc. ICASSP 2008*, Las Vegas, Nevada, USA, April 2008. IEEE.