INTERSPEECH 2009 BRIGHTON

Data-driven Clustering in Emotional Space for Affect Recognition Using Discriminatively Trained LSTM Networks

Martin Wöllmer[1], *Florian Eyben*[1], *Björn Schuller*[1], *Ellen Douglas-Cowie*[2], *Roddy Cowie*[2]

[1]Technische Universität München, Institute for Human-Machine Communication, 80290 München, Germany
[2]Queen's University, School of Psychology, Belfast, BT7 1NN, UK
[woellmer,eyben,schuller]@tum.de

10.21437/Interspeech.2009-474

Abstract

In today's affective databases speech turns are often labelled on a continuous scale for emotional dimensions such as *valence* or *arousal* to better express the diversity of human affect. However, applications like virtual agents usually map the detected emotional user state to rough classes in order to reduce the multiplicity of emotion dependent system responses. Since these classes often do not optimally reflect emotions that typically occur in a given application, this paper investigates data-driven clustering of emotional space to find class divisions that better match the training data and the area of application. Thereby we consider the Belfast Sensitive Artificial Listener database and TV talkshow data from the *VAM* corpus. We show that a discriminatively trained Long Short-Term Memory (LSTM) recurrent neural net that explicitly learns clusters in emotional space and additionally models context information outperforms both, Support Vector Machines and a Regression-LSTM net.
Index Terms: emotion recognition, affective databases, long short-term memory

1. Introduction

Up to now, annotators of databases that are used to train and evaluate emotion recognition engines either focus on assigning discrete classes like *anger*, *happiness*, or *neutral* to the emotionally coloured speech turns [1, 2] or they try to use continuous scales for predefined emotional dimensions such as *valence*, *arousal*, or *dominance* [3, 4]. Yet, both strategies are suboptimal: in the first case the class division has to be determined in advance e.g. by defining emotional prototypes that typically occur in a given database. This implies inflexible, fixed classes that can only be changed by combining or splitting certain classes to reduce or increase the 'emotional granularity' [5]. Annotating and modelling emotional dimensions is more flexible and precise since annotation tools like FEELtrace [6] enable a quasi-infinite resolution of human affect. Yet, when evaluating and processing the output of emotion recognisers that provide continuous values for valence, arousal, etc., the emotional continuum has to be discretised again, e.g. in order to reduce the multiplicity of possible system responses of an emotionally sensitive virtual agent. A common pactice is to use a mapping to *quadrants* such as *positive-active*, *positive-passive*, *negative-active*, and *negative-passive* [7]. However, these classes often do not optimally reflect typical emotional states that occur within the training data or are to be expected when applying the emotion recognition engine in a real-world scenario. For example in [8], the *positive-passive* quadrant had to be excluded since it did not occur in the training set. This suggests that a categorisation of affective sates in the valence-arousal space should not just involve a simple discretisation of the axes but rather closely investigate continuous annotations of the training examples to find meaningful classes.

In this paper we apply a data-driven clustering of the valence-arousal space in order to find classes that better fit the data our recogniser is trained on, and to optimally model the affective states that actually occur in the specific recognition task. Between two and six emotional states are determined via k-means clustering of the training data. Thereby we consider two databases with completely different distributions in emotional space: the Belfast Sensitive Artificial Listener (SAL) database [3] where the occurrence of positive and negative emotions is relatively balanced, and TV talkshow data from the *Vera am Mittag* (VAM) corpus [4] which contains mainly negative emotions. For emotion recognition, both databases imply the great challenge of having to deal with all data - as observed and recorded - and not only with manually selected 'emotional prototypes' as in many other databases.

Further, we introduce a novel emotion recognition strategy that is optimally suited to distinguish between the determined emotional clusters and models *emotional history* by including long-range temporal dependencies into the recognition process. Our technique applies Long Short-Term Memory (LSTM) recurrent neural nets that are explicitly trained on the clusters in a valence-arousal space. By jointly modelling the emotional dimensions, we show that our approach outperforms the LSTM used for regression as introduced in [9]. Furthermore, our experiments reveal that our discriminative LSTM prevails over standard Support Vector Machines (SVM) that are trained on the same task but do not model contextual information.

The structure of this paper is as follows: Section 2 introduces the emotional speech databases that are used in this work, Section 3 gives an overview over the features that are extracted from the speech signal, and Section 4 outlines the principle of an LSTM recurrent neural net. Finally, experimental results are given in Section 5 before concluding in Section 6.

2. Databases

2.1. SAL

The first database we used is the Belfast Sensitive Artificial Listener data which is part of the final HUMAINE database [3]. We considered a subset which contains 25 recordings in total from four speakers (two male, two 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. Data has been labelled continuously in real time by four annotators with respect to valence and activation using a system based on FEELtrace [6]. The adjusted values for valence and activation were sampled every 10 ms to obtain a temporal quasi-continuum. As continuous ground truth label we used the mean of the four annotators.

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 results are not speaker independent, which in turn would not be feasible with only four speakers. Labels for each turn were computed by averaging the frame level valence and activation labels over the complete turn.

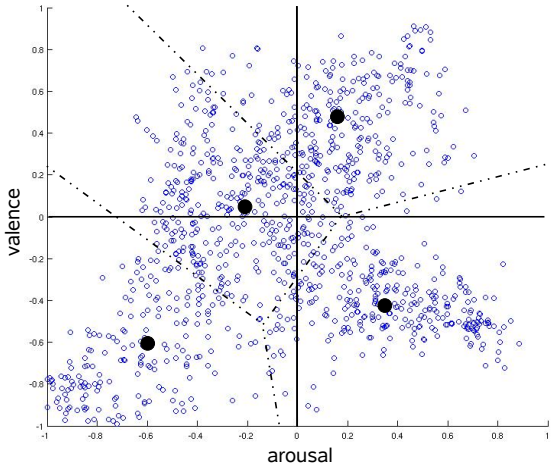

Figure 1: *Annotations of the speech turns in the SAL database with cluster midpoints and class borders (dashed lines) determined via k-means clustering*

Finally, k-means clustering (with Euclidean distance) was conducted to find between two and six clusters and the corresponding class borders in a two-dimensional valence-arousal space. Figure 2 shows the cluster midpoints obtained for four clusters (black points) as well as the annotations of all utterances in the training set in terms of small circles. While three clusters roughly correspond to the common quadrants, one cluster centre marks an emotional state of neutral valence and slightly negative arousal which can hardly be assigned to one of the quadrants but obviously represents a typical affective user state when interacting with virtual agents.

2.2. VAM

The second emotional speech corpus used in this paper is the VAM database [4]. It contains 947 spontaneous and emotionally coloured utterances from 47 guests of the German talkshow "Vera am Mittag" and was recorded from unscripted, authentic discussions. For speaker independent evaluation we randomly selected ten speakers for testing while utterances from the remaining 37 speakers were used as training set. A large number of labellers was used to obtain continuous transcriptions for the emotional dimensions valence, arousal, and dominance (17 labellers for one half of the data, six for the other). In our experiments we omitted *dominance* as further dimension since we found that arousal and dominance are correlated with a high correlation coefficient of 0.9. This limits the gain of additionally incorporating the dominance dimension.

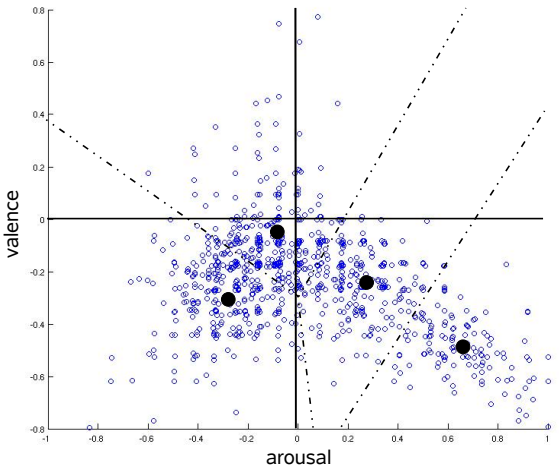

Figure 2: *Annotations of the speech turns in the VAM database with cluster midpoints and class borders (dashed lines) determined via k-means clustering*

As can be seen in Figure 2, due to the topics discussed in the talkshow (friendship crises, defalcation, etc.) mostly negative emotions occur in the database. This points out the need to determine emotional clusters that are representative for affective states occurring within the database. Of course we cannot expect an emotion recognition or automatic TV-show annotation system trained on the valence dimension of VAM data to reliably detect utterances of positive valence, since such speech turns hardly occur in the corpus. In the case of four clusters, all cluster midpoints represent negative valence (see Figure 2).

Feature Group	**Features in Group**	**No.**
Signal energy	Root Mean-Square and log. energy	2
Pitch	Fundamental Frequency F_0, 2 measures for probability of voicing	3
Voice Quality	Harmonics-To-Noise Ratio	1
Cepstral	MFCC	16
Time Signal	Zero-Crossing-Rate, max. and min. value, DC component	4
Spectral	Energy in bands 0-250Hz, 0-650Hz, 250-650Hz, 1000-4000Hz	4
	10%, 25%, 50%, 75%, and 90% Roll-Off	5
	Centroid, Flux, and relative position of maximum and minimum	4
	SUM:	**39**

Table 1: 39 acoustic low-level-descriptors

3. Features

Table 1 lists the 39 acoustic low-level-descriptors that were extracted from the audio signal to train and evaluate our emotion

recognition system. Additionally, first and second order temporal derivatives were used, resulting in 117 features. 51 statistical functionals such as maximum, minimum, mean, quartiles, percentiles, centroids, etc. had been applied, so that the total set consists of 5 967 features. To reduce the feature space dimensionality, relevant features were determined via Correlation based Feature Subset (CFS) selection. Furthermore, all features were normalised to have zero mean and unit variance.

4. Long Short-Term Memory

In order to optimally distinguish the emotional classes obtained through clustering as described in Section 2, we applied discriminatively trained Long Short-Term Memory recurrent neural nets. They were originally introduced by [10] who found that long time lags are inaccessible to existing recurrent neural networks (RNNs) since the backpropagated error either blows up or decays over time (vanishing gradient problem). An LSTM layer is composed of recurrently connected memory blocks, each of which contains one or more recurrently connected memory cells, along with three multiplicative 'gate' units: the input, output, and forget gates. The gates perform functions analogous to read, write, and reset operations. More specifically, the cell input is multiplied by the activation of the input gate, the cell output by that of the output gate, and the previous cell values by the forget gate (see Figure 3). Their effect is to allow the network to store and retrieve information over long periods of time. If, for example the input gate remains closed, the activation of the cell will not be overwritten by new inputs and can therefore be made available to the net much later in the sequence by opening the output gate. This principle overcomes the vanishing gradient problem and gives access to long range context information.

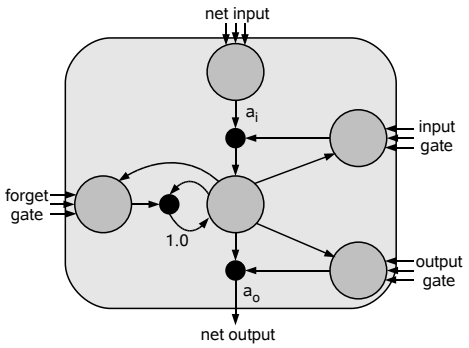

Figure 3: *LSTM memory block consisting of one memory cell: input, output, and forget gate collect activations from inside and outside the block which control the cell through multiplicative units (depicted as small circles); input, output, and forget gate scale input, output, and internal state respectively; a_i and a_o denote activation functions; the recurrent connection of fixed weight 1.0 maintains the internal state*

LSTM networks have demonstrated excellent performance in many tasks that profit from context modelling, e.g. phoneme recognition [11] or keyword spotting [12]. In [9] a regression technique was used to train LSTM networks for the prediction of continuous values for valence and arousal under consideration of *emotional history* (those networks will be referred to as Regression-LSTMs in the following). In this work, however, we discriminatively train an LSTM on the discrete clusters in a way that the size of the output layer corresponds to the number of different emotional clusters that shall be distinguished. For a given speech turn, the activations of the network outputs indicate the probability of the corresponding cluster. Similar to [9], the size of the input layer is equal to the number of acoustic features. In our experiments the hidden LSTM layer contained 100 memory blocks of one cell, each. To improve generalisation, zero mean Gaussian noise with standard deviation 0.6 was added to the inputs during training. The networks were trained using Resilient Propagation (rProp).

5. Experiments

For both databases we evaluated the performance of our discriminatively trained LSTM, the Regression-LSTM as used in [9], and Support Vector Machines on six different emotion recognition tasks: the distinction of two to six emotional clusters as well as the assignment to one of the four quadrants in the valence-arousal space. In contrast to the discriminative LSTM and SVM, the Regression-LSTM outputs continuous values for valence and arousal which were discretised afterwards, according to the clusters and quadrants they would have been assigned to using the minimum Euclidean distance. In order to be able to carry out feature selection separately for valence and arousal, two spearate networks (one for valence and one for arousal) had been trained for Regression-LSTM-based emotion recognition while for the discriminative LSTM and for SVM only one classifier had been trained directly on the discrete cluster or quadrant indices to jointly classify valence and arousal.

Cluster	**2**	**3**	**4**	**5**	**6**	**4(q)**
features	109	132	125	111	102	129
$\mathbf{LSTM_d}$						
accuracy	77.1	61.0	50.7	41.4	40.0	50.5
recall	67.1	55.5	46.4	40.1	37.5	48.1
precision	77.1	59.5	44.6	36.3	35.2	51.6
F1-measure	**71.7**	**57.4**	**45.5**	**38.1**	**36.3**	**49.8**
$\mathbf{LSTM_r}$						
accuracy	70.8	47.1	30.9	38.0	27.5	34.9
recall	58.9	48.6	33.4	33.0	27.8	58.9
precision	64.3	50.0	31.0	34.5	24.3	35.4
F1-measure	**61.5**	**49.3**	**32.2**	**33.8**	**26.0**	**35.6**
SVM						
accuracy	66.1	51.4	38.6	30.0	27.1	41.4
recall	55.3	46.6	38.1	30.3	26.0	41.4
precision	57.6	43.7	34.6	27.9	23.7	42.2
F1-measure	**54.9**	**42.0**	**32.8**	**25.2**	**21.8**	**38.9**
dummy						
accuracy	68.3	60.2	44.1	31.7	30.7	35.9
recall	50.0	33.3	25.0	20.0	16.7	25.0

Table 2: *SAL database: number of selected features and results for the discrimination of 2, 3, 4, 5, and 6 emotional clusters as well as for the 4 quadrants (4q) when using discriminatively trained LSTM ($LSTM_d$), Regression-LSTM ($LSTM_r$), Support Vector Machines (SVM), or a 'dummy' feature (for chance reference)*

Table 2 shows the performance of the different classifiers for six different recognition tasks using the SAL database. For chance reference, the results obtained through a single constant 'dummy' feature resulting in picking the majority class at any time are included. Note that due to unbalanced class distributions, accuracy is a rather inappropriate performance measure. Thus, we used the F1-measure as the harmonic mean between

unweighted recall and unweighted precision for performance comparison. As can be seen, the discriminative LSTM outperforms both, the Regression-LSTM and SVM. Since in the SAL database all quadrants are sufficiently 'occupied' (see Figure 1), the F1-measure for the discrimination of four quadrants is slightly higher than for the discrimination of four emotional clusters. However, this is not true for the VAM corpus (see Table 3). There, two quadrants are almost unoccupied (see Figure 2), which leads to better F1-measures for the discrimination of four clusters and highlights the importance of defining class borders according to the application and the database respectively rather than just discretising emotional space to equidistant fields.

Cluster	**2**	**3**	**4**	**5**	**6**	**4(q)**
features	155	150	141	145	132	140
$\mathbf{LSTM_d}$						
accuracy	82.1	71.3	59.0	45.6	48.2	74.4
recall	80.7	75.8	63.0	50.3	47.4	41.3
precision	75.8	69.2	59.5	47.6	47.6	36.8
F1-measure	**78.2**	**72.3**	**61.2**	**48.7**	**47.5**	**38.9**
$\mathbf{LSTM_r}$						
accuracy	85.6	72.3	52.8	43.1	43.6	67.2
recall	80.8	71.5	55.5	45.9	41.3	38.8
precision	80.0	71.4	57.8	49.2	32.7	42.5
F1-measure	**80.4**	**71.5**	**56.6**	**47.5**	**36.5**	**40.6**
SVM						
accuracy	81.5	68.7	53.8	46.2	45.1	71.8
recall	75.1	70.5	56.8	50.1	45.0	41.1
precision	74.4	67.6	56.0	49.2	43.2	48.1
F1-measure	**74.7**	**68.9**	**56.1**	**47.9**	**43.3**	**40.1**
dummy						
accuracy	76.4	51.8	43.1	28.2	33.9	52.3
recall	50.0	33.3	25.0	20.0	16.7	25.0

Table 3: *VAM database: number of selected features and results for the discrimination of 2, 3, 4, 5, and 6 emotional clusters as well as for the 4 quadrants (4q) when using discriminatively trained LSTM ($LSTM_d$), Regression-LSTM ($LSTM_r$), Support Vector Machines (SVM), or a 'dummy' feature (for chance reference)*

Apart from the quadrant discrimination and the task of distinguishing two clusters in the VAM corpus, the discriminative LSTM again prevails over the Regression-LSTM and the SVM.

On both datasets, the absolute F1-measure is rather low compared to results for the discrimination of 'prototypical emotions' as published in [13], for example. Yet, in real-life applications of emotion recognition, not only unambiguous emotions have to be classified. The challenge for next-generation emotion recognition systems is rather to develop advanced classifiers using long-range context to continuously deal with all data - as it is necessary for the databases used herein.

6. Conclusion

In this work we investigated data-driven clustering of the valence-arousal space as an alternative to quadrant-based quantisation that ignores typical emotions occurring in a given scenario. Furthermore, we designed a novel discriminative LSTM network which exploits long-range context information and outperforms conventional Support Vector Machines in terms of emotional cluster classification performance.

In future works we will focus on *emotion detection* in order to detect sudden strong emotions within long sequences of neutral affect. Moreover, we will investigate the benefit of updating emotion prediction using *bidirectional* LSTM networks.

7. Acknowledgements

We would like to thank Alex Graves for his support concerning LSTM networks.
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] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. of Interspeech*, Lisbon, Portugal, 2005, pp. 1517–1520.

[2] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russel, and M. Wong, ""you stupid tin box" - children interacting with the aibo robot: a cross-linguistic emotional speech corpus," in *Proc. of LREC*, Lisbon, Portugal, 2004, pp. 171–174.

[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*. Lisbon, Portugal: Springer-Verlag Berlin Heidelberg, 2007, vol. 4738, pp. 488–500.

[4] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *Proc. of ICME*, Hannover, Germany, 2008, pp. 865–868.

[5] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application," *Image and Vision Computing Journal (IMAVIS), Special Issue on Visual and Multimodal Analysis of Human Spontaneous Behavior*, p. 17 pages, 2009, to appear.

[6] 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 *Proc. of the ISCA Workshop on Speech and Emotion*, 2000, pp. 19–24.

[7] R. Plutchik, *Emotion: A psychoevolutionary synthesis*. NY, USA: Harper and Row, 1980.

[8] R. Cowie, E. Douglas-Cowie, J. G. Taylor, S. Ioannou, M. Wallace, and S. Kollias, "An intelligent system for facial emotion recognition," in *Proc. of ICME*, 2005.

[9] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. of Interspeech*, Brisbane, Australia, 2008, pp. 597–600.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9(8), pp. 1735–1780, 1997.

[11] A. Graves, S. Fernandez, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *Proc. of ICANN*, Warsaw, Poland, 2005, pp. 602–610.

[12] M. Wöllmer, F. Eyben, J. Keshet, A. Graves, B. Schuller, and G. Rigoll, "Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks," in *Proc. of ICASSP*, Taipei, Taiwan, 2009.

[13] B. Schuller, M. Wimmer, L. Mösenlechner, D. Arsic, and G. Rigoll, "Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space?" in *Proc. of ICASSP*, Las Vegas, NV, 2008, pp. 4501–4504.