A Multitask Approach to Continuous Five-Dimensional Affect Sensing in Natural Speech

FLORIAN EYBEN, MARTIN WÖLLMER, and BJÖRN SCHULLER, Institute for Human-Machine Communication, TUM, Germany

Automatic affect recognition is important for the ability of future technical systems to interact with us socially in an intelligent way by understanding our current affective state. In recent years there has been a shift in the field of affect recognition from "in the lab" experiments with acted data to "in the wild" experiments with spontaneous and naturalistic data. Two major issues thereby are the proper segmentation of the input and adequate description and modeling of affective states. The first issue is crucial for responsive, real-time systems such as virtual agents and robots, where the latency of the analysis must be as small as possible. To address this issue we introduce a novel method of incremental segmentation to be used in combination with supra-segmental modeling. For modeling of continuous affective states we use Long Short-Term Memory Recurrent Neural Networks, with which we can show an improvement in performance over standard recurrent neural networks and feed-forward neural networks as well as Support Vector Regression. For experiments we use the SEMAINE database, which contains recordings of spontaneous and natural human to Wizard-of-Oz conversations. The recordings are annotated continuously in time and magnitude with FeelTrace for five affective dimensions, namely activation, expectation, intensity, power/dominance, and valence. To exploit dependencies between the five affective dimensions we investigate multitask learning of all five dimensions augmented with inter-rater standard deviation. We can show improvements for multitask over single-task modeling. Correlation coefficients of up to 0.81 are obtained for the activation dimension and up to 0.58 for the valence dimension. The performance for the remaining dimensions were found to be in between that for activation and valence.

Categories and Subject Descriptors: H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*; H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing—*Modeling*

General Terms: Algorithms, Experimentation, Human Factors

Additional Key Words and Phrases: Neural networks, long short-term memory, emotion recognition, audio features, SEMAINE, dimensional affect

ACM Reference Format:
Eyben, F., Wöllmer, M., and Schuller, B. 2012. A multitask approach to continuous five-dimensional affect sensing in natural speech. ACM Trans. Interact. Intell. Syst. 2, 1, Article 6 (March 2012), 29 pages.
DOI = 10.1145/2133366.2133372 http://doi.acm.org/10.1145/2133366.2133372

1. INTRODUCTION

As the number of technical gadgets and electronic devices, which play a role in our everyday lives, constantly grows, intuitive and easy interaction becomes more and more an essential factor. The way we interact with computers, service machines, and

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).
Authors' addresses: F. Eyben (corresponding author), M. Wöllmer, and B. Schuller, Institute for Human-Machine Communication, TUM, Germany; email: eyben@tum.de.

DOI 10.1145/2133366.2133372 http://doi.acm.org/10.1145/2133366.2133372

household appliances, for example, is far from being as convenient as interacting with fellow humans. We spend unnecessary time by adapting to different user interfaces and learning how to control devices. Often, malfunction or nonobvious functionality leads to anger and frustration of users. Many people, especially elder ones, thus are afraid of using modern computer technology.

A key in achieving faster, more intuitive interaction is to make machines understand our intentions in a similar fashion as our human peers do. That is, there is a need for socially intelligible machines, which robustly accept multimodal and sometimes even ambiguous input, and deduce the user's intention based on, for example, his/her verbal and nonverbal behavior, affective state, situational context, and background knowledge. This article focuses on detecting the user's affective state. Affect plays a major role in human-machine interactions since it can be a very reliable indicator for inappropriate machine responses and wrong-goings in the interaction, for example. Moreover, being aware of the user's affective state to a certain degree can enable virtual agents and service robots to react more appropriately to the current situation.

While Automatic Emotion Recognition (AER) from acted, emotionally prototypical read speech gives results comparable to human performance (refer to Burkhardt et al. [2005] and Schuller et al. [2009c]) and thus seems to be solved, reliable affect recognition in natural, changing environments from spontaneous—and maybe also nonacted—speech, in contrast, remains a challenge at present [Schuller et al. 2009b, 2010a]. There are numerous reasons why automatic recognition of spontaneous emotions poses such a challenge. First, the spoken content in the natural utterances is not fixed, which makes it harder, if not impossible, to train word-dependent emotion models as can be done for corpora like the Berlin Speech Emotion Database (EMO-DB) [Burkhardt et al. 2005], for example. Next, the full continuum of possibly expressible emotions can occur and must be discriminated, that is, fine-grained differences between often very similar and subjective nonprototypical emotions must be handled. This cannot be done by categorical modeling of emotions, instead an approach for continuous representation of affect in a dimensional space is usually chosen [Cowie et al. 2000; Douglas-Cowie et al. 2007a; Grimm et al. 2007a; Wöllmer et al. 2008]. Different approaches which try to mitigate the problmes of both categorical and dimensional approaches are presented in Mower et al. [2011] and Mower and Narayanan [2011]. Both the categorical labeling approach and the dimensional approach suffer from the partly subjective nature of affect, that is, large inter-subject variations in perceived affect type and strength, which leads to moderate or low inter-labeler agreement. In the past, segments with low inter-labeler agreement were removed, yielding only prototypical emotions, which fitted the categories well. For spontaneous affective speech, this approach seems not feasible, since a system "in the wild" has to deal exactly with these ambiguous cases.

Another important issue for AER in natural environments is the segmentation of the input. Especially for real-time interactive systems this is a crucial issue. AER mostly deals with recognizing emotion from large units of speech [Zeng et al. 2009], for example, complete sentences or fragments of sentences [Schuller et al. 2006; Wöllmer et al. 2008]. The fragments in most cases are presegmented and all results obtained have the precondition of perfect segmentation. In reality perfect segmentation is not possible. Moreover, the segments are quite long, which adds a considerable latency to the recognition system, since the complete segment must be recorded before it can be analysed and a prediction can be produced. Studies on the influence of the unit length on recognition performance have been conducted in Batliner et al. [2010] and Mower and Narayanan [2011]. The shortest feasible unit of analysis used was the word level. To obtain a perfect segmentation in a live system for the word level is near to impossible and requires a full-blown ASR system running and consuming computational resources. Thus, we will investigate alternate methods of segmentation

which are invariant to segmentation errors, consider a sufficient amount of context, and can be adapted to output emotions at any given rate.

All those aspects highlight how important it is to move forward in the field of robust incremental affect recognition in natural environments. In this article we go beyond the simple evaluation of features at the end of a speech turn as it is applied in various conversational systems (e.g., Streit et al. [2006]). We modify the existing turn-based approach to an incremental supra-segmental approach (see Section 2.3 for details on the supra-segmental approach), and use a model (Long Short-Term Memory Recurrent Neural Networks, LSTM-RNN) that is able to utilize long-range dependencies between consecutive segments. With the same network we move another step forward and evaluate and discuss the potential of this method to predict trace-style dimensional affect labels directly from low-level audio feature frames. In a multitask learning setup we estimate the confidence of the automatic predictions by training the networks with inter-rater standard deviations as additional target, which also brings a mutual benefit to the original single target. Further, we evaluate a novel method for multidimensional affect recognition: a multitask learning setup where five affective dimensions and the corresponding inter-rater standard deviations are modeled by a single network. In contrast to previous work [Wöllmer et al. 2008; Eyben et al. 2010a], where only two affective dimensions, namely, activation and valence, were used, we herein evaluate performance for three additional affective dimensions, namely expectation, intensity, and power. The use of these dimensions is motivated in more detail in Section 2.2.

The remainder of this article is structured as follows: In the next section (2) we give a more extensive overview on related work and challenges of automatic recognition of natural, spontaneous affect from speech audio input. A large-scale database of spontaneous affective interactions, which has been recorded in the course of the SEMAINE project, is used for evaluations in this study. It is described in Section 3. Our proposed methods towards robust, low-latency, continuous multidimensional affect recognition, which are based on multitask learning with Long Short-Term Memory Recurrent Neural Networks, are introduced and described in Section 4. The obtained evaluation results are discussed in Section 5, and conclusions are drawn in Section 6.

2. STATE-OF-THE-ART AND RELATED WORK

In the Introduction three major issues were mentioned which are crucial for emotion recognition in real-world deployable applications. Related work and current approaches to these issues will be discussed throughout this section. Section 2.1 contrasts emotion recognition experiments performed under highly restricted conditions with those performed on real-world data, and highlights the challenges of the latter case. The current state-of-the art in dimensional affect recognition is summarized in Section 2.2, and issues concerning the trade-off between the analysis segment length and the accuracy and latency in an online system are discussed in Section 2.3.

2.1. Affect: "in the Wild" vs. "in the Lab"

It is often believed that emotion recognition from speech is solved because numerous publications in the past have reported high accuracies (above 80%) for Ekman's basic six emotions [Ekman and Friesen 1975], for example. Most of these have analysed read speech, where prototypical emotion categories were acted out. The most well-known and widely used such dataset is the Berlin Speech Emotion database (EMO-DB) [Burkhardt et al. 2005]. Ambiguous sentences, with low interlabeler agreement, were removed from the dataset resulting only in prototypical samples. These samples can be identified with high accuracy with models trained on different data from the same corpus [Vlasenko and Wendemuth 2007; Schuller et al. 2009c]. When performing

cross-corpus experiments, even between corpora with acted and prototypical emotions, accuracies drop significantly [Schuller et al. 2010c]. This shows that the models are very corpus-specific or even specific to the linguistic content. In the case of EMO-DB, for example, only a small set of different sentences are spoken with different emotions. The same sentences occur in test and training splits. Another reason for the performance drop might be that practically no two corpora with exactly the same categories exist. In Schuller et al. [2010c] similar categories are mapped to related categories or a binary grouping into positive and negative valence and high/low arousal. In this light, the dimensional representation seems to be more universal when it comes to cross-corpus and cross-domain experiments.

Recent work has tackled the challenge of automatically identifying natural affect. In Devillers et al. [2005], turns are assigned multiple targets (mixtures of emotion categories) based on a realistic affective speech dataset with nonacted speech. Schuller et al. [2007a] also investigate the issue of emotion recognition on realistic and nonprototypical data. Recent INTERSPEECH challenges have attracted great interest and advances in the field and demonstrated how challenging the matter is [Schuller et al. 2009b, 2010a, 2010b, 2011].

2.2. Affect Representation in a Five-Dimensional Space

Automatic dimensional affect recognition is still in an early stage [Grimm et al. 2007b; Wöllmer et al. 2008; Schuller et al. 2009a; Gunes and Pantic 2010b, 2010a]. The most commonly employed strategy is to reduce the dimensional emotion classification problem to a two-class problem (positive versus negative or active versus passive classification, for example, Nicolaou et al. [2010], Schuller et al. [2009c], and Wöllmer et al. [2010a]), a four-class problem (classification into the quadrants of 2D V-A space, for example, Caridakis et al. [2006], Fragopanagos and Taylor [2005], Glowinski et al. [2008], and Ioannou et al. [2005]), or to automatically identify clusters in the emotional space [Wöllmer et al. 2009, 2010b; Lee et al. 2009]. Introducing fixed clusters or categories brings up the problem of ambiguity again. Instances originally rated with a dimensional label on or near to the cluster boundary are much more likely to be assigned to the wrong cluster in the evaluation step. The results are degraded because during evaluation it is generally not distinguished between confusion of neighboring clusters and confusion between clusters further apart in the dimensional space. A feasible solution is a regression model which directly predicts the continuous values of the dimensions. Only very few works on this topic exist so far: For example, Grimm et al. [2007a] use Support Vector Regression to predict affect in three dimensions (activation, valence, power/dominance), Wu et al. [2010a] attempt fusion of three methods: robust regression, Support Vector Regression, and locally linear reconstruction, Wöllmer et al. [2008] use Long Short-Term Memory Neural Networks and Support Vector Machines for Regression (SVR), and the work presented in Wöllmer et al. [2010b] utilizes a Bidirectional Long Short-Term Memory Neural Networks performing regression for emotion dimensions and quantizing the results into four quadrants (after training). Our previous work in Eyben et al. [2010a] also investigates a regression technique for continuous dimensional affect recognition. Alternative methods to Support Vector Regression include linear regression [Cohen et al. 2003], radial base function networks [Yee and Haykin 2001], or standard feed-forward perceptron networks (standard neural networks). In the context of emotion-related virtual agents Recurrent Neural Networks with Long Short-Term Memory [Hochreiter and Schmidhuber 1997] have been suggested [Peters and O'Sullivan 2002; Eyben et al. 2010a]. An approach incorporating body language for recognition of continuous emotion states is reported in Metallinou et al. [2011].

However, all the previous approaches have reported results on no more than three affect dimensions. With the availability of the SEMAINE database (refer to Section 3), experiments with two new and unexplored dimensions, expectation and intensity, are possible. The choice of these dimensions is explained in McKeown et al. [2010]. It is based on psychological findings reported in Fontaine et al. [2007]: There, the dimensions activation, valence, expectation, and power were obtained by a Principal Component Analysis (PCA) applied to 144 hand-assigned "emotion features" derived from terms people use to describe emotional events. Activation indicates the level of arousal, that is, the level of active engagement or readiness for action, versus passiveness as found in contentment or boredom, for example. On the valence dimension pleasant versus unpleasant emotions are contrasted, that is, valence is an indication of how positive (pleasant) or negative (unpleasant) the emotion is. The dimension "power" characterizes whether the emotion is related to a feeling of power and control or weakness. Pride and anger are opposed to sadness and despair, for example. Expectation is a measure of unpredictability versus expectedness or familiarity. Surprise, fear, and disgust, for example, are thus characterized by a low expectation value, while all other emotions, such as stress and contempt, are associated with a higher expectation value as they occur in contexts more familiar to the subject. The fifth dimension (intensity) was added by the creators of the database as an overall measure of perceived emotional intensity, that is, the distance of the current sample from the center of general neutrality, regardless on which dimension.

The only other database so far, as known to the authors, that contains more than three annotated dimensions is the CINEMO database [Schuller et al. 2010d]. One of the first publications reporting on experiments with all of these five dimensions on the SEMAINE database is Gunes and Pantic [2010c], which focuses on dimensional prediction of emotions from spontaneous conversational head gestures by mapping the amount and direction of head motion, and occurrences of head nods and shakes into activation, expectation, intensity, power, and valence level of the observed subject using Support Vector Regression. All existing multidimensional affect recognition approaches use a separate model for each dimension. However, dimensions are often correlated to some extent (e.g., activation and intensity), thus a joint modeling might boost performance.

When moving away from categorical affect recognition and classification methods towards regression analysis of dimensional affect, we at first lose an important output measure: the classifier's confidence. Although Support Vector Regression models as in Chang and Lin [2001] do support probability estimates, this information is of limited practical use. So far, in closely related affect recognition literature (to the best of our knowledge) no experiments on confidence estimation of regression predictions have been reported. A technique for dimensional music mood prediction has been introduced in Schmidt and Kim [2012]. The authors use linear regression to estimate the mood coordinates of a song excerpt in a 2D activation-valence space and the uncertainty is thereby modeled as an additional regression target. For training the system the authors collected a continuously annotated database through an online game in which participants had to label the current mood of a song on an activation valence map. They thereby competed against other players. Those players whose labels were most similar to their opponent's labels were awarded the highest score. In this article we propose a similar attempt to estimate the confidence for speech affect by multitask learning with neural networks as regressors. The networks thereby model the human inter-rater standard deviation of the training data along with the mean label. The idea of using multitask learning also promises to boost performance for the main task. The multitask approach is also inspired by the work presented in Steidl et al. [2009], where a similar technique is employed for estimating class confidences.

2.3. The Unit of Analysis

As opposed to speech recognition, emotion recognition from isolated short-time audio frames is virtually impossible: While single phonemes are highly correlated to a specific spectral representation in short signal windows, speech emotion is a phenomenon observed over a longer time window (typically more than 1–2 seconds). Typical units of analysis are complete sentences, sentence fragments (i.e., chunks, e.g., by syntactical rules) or words [Steidl 2009]. The term "segment" will be used in the ongoing referring to a general unit of analysis. Finding the optimal unit of analysis is still an active area of research [Schuller and Rigoll 2006; Schuller et al. 2007b, 2008; Busso et al. 2007; Mower and Narayanan 2011]. As stated in Zeng et al. [2009], the segmentation is one of the most important issues for real applications but has been "largely unexplored so far". An in-depth study on the effect of the analysis unit length can be found in Batliner et al. [2010].

Traditional audio feature extraction approaches are based on short-time spectral analysis, where windows of typically 25 ms, in which the signal is assumed more or less stationary, are used as low-level analysis frames. Features on this level are referred to as low-level features or Low-Level Descriptors (LLD). Classifying low-level feature vectors directly and independently with respect to their emotional content is not feasible, since emotion is mainly expressed by the evolution of these features over a certain time period (e. g., prosody!). Thus, a context spanning multiple feature vectors must be considered. To do so, the most widespread method is the mapping of the sequence of LLD belonging to a segment to a single high-dimensional vector by applying statistical functionals, such as mean and moment. This technique is referred to as supra-segmental modeling. It enables mapping of sequences of variable length to a vector of fixed dimensionality. Both classification (for affect classes) and regression (affect dimensions) tasks can be solved by this approach, given suitable modeling, for example, Support Vector Machines (SVM) for classification and Support Vector Regression (SVR) for regression. A major drawback of these approaches is that one complete input fragment is required for analysis and only a single output can be produced at the end of every input fragment, which is typically a sentence or part of sentence. Thus, true continuous output at a fixed rate in the second or sub-second region is not possible with this approach, except by interpolation of the output from higher levels, which gives no new information. Alternative approaches do not model the long-range dependencies on the feature level but instead use hidden Markov modeling. As feature vectors the low-level descriptor frames are used (refer to, e.g., Schuller et al. [2003] and Vlasenko et al. [2007]). Yet, these approaches also require the complete input fragment at hand to perform a best-path decoding. Moreover, they can only produce one discrete class output per fragment and are therefore unsuited for dimensional emotion recognition.

For fully continuous emotion recognition we must ideally abandon the requirement of defining a suitable unit of analysis, within which the emotional state is assumed as quasi-stationary. Under ideal circumstances, only frame-wise features should be used, the long-range dependencies must be modeled by the classifier/regressor, and it should be possible to obtain an output of the current state for every input frame. In Section 4.1 we will present a classifier which meets all these requirements. We evaluate to what extent such an approach is feasible, or whether a modified supra-segmental approach is better.

3. THE SEMAINE DATABASE

The SEMAINE database [McKeown et al. 2010] was recorded to study natural social interaction that occurs in conversations between humans and the future generation of artificially intelligent agents, and to collect training data for such intelligent agents,

especially the SEMAINE system. The database is freely available for scientific research from `http://semaine-db.eu`. The scenario used for provoking emotionally colored, naturalistic interactions is the Sensitive Artificial Listener (SAL) scenario. It involves a user interacting with emotionally stereotyped "characters" whose responses are stock phrases provoked by the user's emotional state rather than the content of what he/she says. The model is a style of interaction observed in chat shows and parties, which aroused interest because it seems possible that a machine with some basic emotional and conversational competence could sustain such a conversation, without needing to be competent with fluent speech and language understanding.

In the recording scenario, the participants are asked to talk to each of the four emotionally stereotyped characters in turn. These are Prudence, who is even-tempered and sensible, Poppy, who is happy and always outgoing, Spike, who is angry and confrontational, and Obadiah, who is depressive and sad. The study presented in this work is based on the first part of the SEMAINE database, the Wizard-of-Oz part. In this part, human operators pretended to be the artificial agents. This type of interaction is called Solid-SAL. Because we assume that an automatic SAL agent has no language understanding, a few rules govern this type of interaction. The most important of these is that the agent is not allowed to answer questions. However, the operators are instructed that the most important aspect of their task is to create a natural style of conversation; strict adherence to the rules of a SAL engagement was secondary to a conversational style that would produce a rich set of conversation-related behaviors and therefore transgressions occasionally occur, however, only very rarly (roughly less than 1–2% of sentences), most of the time due to subjects asking questions, and the operator answering them.

Audiovisual recordings of the full Solid-SAL interactions exist of both the user and the operator, each with a color and greyscale frontal-view camera and an additional side-view camera for the user. Collar and table microphone recordings were conducted for both user and operator. The audio was recorded at 48 kHz with 24 bits per sample. For research in audio-visual fusion on the feature level, the audio and video signals were synchronized with an accuracy of 25 μs using the system developed by Lichtenauer et al. [2010]. For this study we use only the audio portions, specifically the user's speech turns, of the Solid-SAL part of the database.

The Solid-SAL part of the database holds recordings of 20 trials of the SAL experiment, split into over 100 character conversations of approximately 5 minutes each. All recorded conversations have been fully transcribed and annotated for five affective dimensions and partially annotated for 27 other dimensions, using trace-style continuous ratings (similar to FeelTrace [Cowie et al. 2000]). Thereby the annotators could move a slider continuously in a given range while listening to the recording in order to rate their current opinion regarding a single affective dimension at a time. The ratings from the slider were sampled at a rate of 50 per second and with a granularity of 0.001. The five core dimensions are those that psychological evidence suggests are best suited to capture affective coloring in general [Fontaine et al. 2007]. They are *valence*, *activation*, *power*, *anticipation/expectation* with the addition of overall emotional *intensity*. We would like to note at this point that the dimension *intensity* in the SEMAINE database appears to be highly correlated with arousal (correlation coefficient of 0.67; see Table V). We still decided to report results on this dimension, as it has been chosen as fith dimension in the SEMAINE project and the annotations are available in the SEMAINE database. However, the reader is to mind the high correlation between these dimensions when interpreting the results. More details on these dimensions are given in Section 2.2. In total, trace-style ratings for all five affective dimensions exist from eight raters. However, at the time of writing not all raters had rated all sessions, thus we chose to include only those sessions in our experiments where the minimum

Table I. Summary of the SEMAINE Database Statistics

Subjects	21 ()
Characters	4 (Poppy, Spike, Obadiah, Prudence)
Recordning Sessions	57
User speech turns	2 189
User speech total time	5 hours

number of raters was three. Moreover, categorical labels were assigned to segments in the database in a tag-like manner. As they are not used in this article, we refer the reader to McKeown et al. [2010] for details.

The raters were all experienced psychology students and were all instructed about the meaning of these dimensions. They were instructed to provide ratings for their overall sense of where an individual at any instant in time should be placed along a given dimension. They did this by watching the video and listening to the audio of one recording session and adjusting the trace slider for one selected dimension accordingly as the video went along. Corrections were not possible. The raters had to watch each video once for every dimension that was annotated. It is important to note that the judgements of the raters are based on the video and audio at the same time. Unfortunately, no separate ratings exist. Therefore we can expect certain regions in the ratings to be more correlated to the audio and others more correlated to the video. Further details on the annotation guidelines can be found in Douglas-Cowie et al. [2007b].

To obtain a single target value for each dimension, the values of the individual raters were averaged. For all evaluations we use this mean label as target, which is referred to as mean label or mean dimension label in the ongoing. In addition to this mean label, the standard deviation of all 3–4 raters is computed as a inter-rater confidence measure. In Wöllmer et al. [2008], we performed a normalization of the labels for each rater before computing the average to compensate for inter-rater scale mismatches and offsets. In contrast to the SAL database, we did not observe large-scale differences and offsets for the SEMAINE database, and thus decided not to normalize the data.

In total there are 20 recordings with 3–4 sessions on average. After sorting out those with two or less raters, 57 sessions remain. From these, 36 sessions are used for training, 14 sessions for evaluation, and 7 sessions as a development set. The sequence IDs of the training sessions as used in the publicly available SEMAINE database are 34–37, 40–43, 46–49, 58–61, 70–73, 76–79, 82–85, 88–91, and 94–97. Those of the development set are 19–22, and 29–31, and those of the evaluation set are 13–16, 25–27, 52, 53, 55, and 64–67. In total there are 2 189 user speech turns. Table I gives a summary of the corpus statistics. We ensured gender balance of the subjects in the evaluation set by including sessions from four subjects in total, two males and two females. The training, development, and evaluation sets are subject disjunctive, that is, data from no subject occurs in more than one set.

The training set contains 1 584 user speech turns, where a turn is defined as a continuous segment of user speech bounded either by initial or final silence or a segment of operator speech. The turns have been manually annotated in the database. The development set contains 169 user speech turns. Table II shows detailed statistics concerning the distribution of the “ground truth” dimensional affect labels for all the user speech turns in the training and development set of the SEMAINE database. The evaluation set contains 436 user speech turns. Table III shows detailed statistics concerning the distribution of the “ground truth” dimensional affect labels for the evaluation set only. The figures roughly correspond to those of the whole corpus, which shows that the data in the test set reflects the overall conditions of the corpus relatively well.

Table II. Statistics of the Dimensional Affect Ratings for the Joint Training, and Development Set of the SEMAINE Corpus as Used in This Article

Dimension:	Min. value	Max. value	Mean μ	Std. deviation σ
A_μ	−0.798	0.656	−0.043	0.222
A_σ	0.000	0.594	0.259	0.102
E_μ	−0.769	0.604	−0.358	0.190
E_σ	0.000	0.720	0.223	0.127
I_μ	−0.899	0.697	−0.144	0.188
I_σ	0.000	0.597	0.254	0.105
P_μ	−0.747	0.749	0.417	0.189
P_σ	0.000	0.664	0.165	0.116
V_μ	−0.965	0.887	0.040	0.320
V_σ	0.000	0.499	0.124	0.075

Five dimensions A(ctivation), E(xpectation), I(ntensity), P(ower), V(alence). Mean of all raters of the mean turn label (μ subscript), inter-rater standard deviation for the mean turn label (σ subscript). Minimum/Maximum value, mean, and standard deviation of μ and σ for each dimension.

Table III. Statistics of the Dimensional Affect Ratings for the *Evaluation* Set of the SEMAINE Corpus Used in this Article

Dimension:	Min. value	Max. value	Mean μ	Std. deviation σ
A_μ	−0.582	0.480	−0.027	0.250
A_σ	0.000	0.547	0.247	0.087
E_μ	−0.730	0.441	−0.351	0.229
E_σ	0.000	0.773	0.254	0.140
I_μ	−0.548	0.648	−0.118	0.270
I_σ	0.000	0.425	0.189	0.080
P_μ	−0.350	0.718	0.339	0.237
P_σ	0.000	0.610	0.220	0.141
V_μ	−0.715	0.659	0.008	0.321
V_σ	0.000	0.401	0.119	0.075

Five dimensions A(ctivation), E(xpectation), I(ntensity), P(ower), V(alence). Mean of all raters of the mean turn label (μ subscript), standard deviation of the raters for the mean turn label (σ subscript). Minimum/Maximum value, mean, and standard deviation of μ and σ for each dimension.

Please note that the minimum values of the inter-rater standard deviation for each dimension appear very close to zero in Tables II and III. In theory this indicates perfect rater agreement at some points throughout the sessions. While this can happen at random, we more likely attribute this to the fact that during the process of rating the trace sliders were often initialized to the same value at the beginning of the session. This remained for a few seconds until the raters decided to move the sliders to a different position. Therefore these low values are supposedly not a good indicator of the actual minimal values of the inter-rater standard deviations.

Additionally, note that the dimension *expectation* was scaled from its original range ($[0; 100]$) according to Eq. (1) in order to be in the same range ($[-1; +1]$) as the other four dimensions.

$$E^* = \frac{E}{50} - 1.0 \tag{1}$$

From Table II we can see that for the training and development set the average inter-rater standard deviation (σ subscript) for each turn is approximately 0.2 with a maximum up to approximately 0.7. This highlights the issue of subjectivity of the problem and the great variance among individual rater opinions, which is far more pronuounced on some sentences than others. Moreover, the numbers show that there

Table IV. Correlation Coefficients (CC) between Raters for Each of the Five Dimensions, Computed on the Evaluation Set Sessions.

Dimension: / Rater:	**R1-R3**	**R1-R5**	**R1-R6**	**R3-R5**	**R3-R6**	**R5-R6**	**Avg**
A	0.655	0.552	0.541	0.667	0.494	0.495	**0.567**
E	0.331	0.116	0.216	0.394	0.233	0.212	**0.250**
I	0.694	0.635	0.401	0.696	0.576	0.535	**0.590**
P	0.306	0.172	0.217	0.452	0.064*	0.156	**0.228**
V	0.756	0.720	0.750	0.790	0.829	0.779	**0.771**

Correlations marked with * are *not* statistically significant on a level of $p = 0.05$ using a 2-tailed test.

Table V. Correlations between the Five Dimensions (CC), Computed on the Evaluation Set Sessions

Dimension:	**E**	**I**	**P**	**V**
A	−0.136	0.673	0.126	−0.125
E		0.132	−0.659	0.004
I			−0.287	−0.496
P				0.220

All correlations, except for valence for dimension E) are significant on a level of $p = 0.05$ based on a 2-tailed test.

are turns with higher rater agreement and turns with substantially lower agreement (higher standard deviation) than the mean agreement.

Another method to asses global inter-rater agreement is to compute correlation coefficients between the rater's labels. Table IV shows a pairwise correlation between four raters (R1,R3,R5,R6). The names of the raters are the same as those used in the SEMAINE database, thus there are jumps in the rater ID numbers. The correlation coefficients are reported for the evaluation set sessions, to enable direct comparison of these results with the automatically obtained results in Section 5. Notably, human agreement is highest for valence and lowest for power (also referred to as dominance). For machine-based recognition it is commonly observed (and also found in this article) that valence is one the most difficult dimensions to predict correctly from acoustic parameters alone. People, in many cases, rely on the meaning of the words in the sentence to asses whether it has positive or negative valence. The acoustic parameters, on the other hand, are very good indicators of arousal and intensity.

Given the fact that the inter-rater correlations for *expectation* and *power* are very low (roughly 0.25), we must question whether they provide a reliable ground truth to train models on. While on the one hand it certainly is not overly reliable, there is still some valid ground as some raters agree far better than others. This might indicate that every rater might have judged the levels of these dimensions based on different acoustic cues. Every one of them could be consistent with itself, though. As the approach presented in this article is in principle capable of modeling the behavior of every single rater, it will be part of follow-up work to analyse the performance when building and evaluating models for every single rater. Moreover, we decided to include the results for the *power* and *valence* dimensions, despite the poor rater agreement on these dimensions, to investigate the automatic classification performance in the multitarget learning and to verify whether the ground truth provides some value or must be regarded as invalid.

To better understand the relations between the five dimensions, Table V shows the inter-dimension correlation coefficients on the evaluation set. The most obvious correlation can be seen between activation and intensity (0.673), which shows that a high overall emotional intensity often occurs together with high arousal of the subject. Next, we see that expectation and power are anti-correlated, that is, a high value of expectation is often associated with a low value power/control. An anti-correlation can

also be observed for intensity and valence, which means that negative emotions are expressed with a higher intensity than positive emotions in the SEMAINE set. The remaining emotion pairs can be considered as uncorrelated.

4. PROPOSED APPROACH

To achieve the goal of incremental affect recognition in real time we chose Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN) as regressors. These neural networks are able to successfully model long-term dependencies between inputs, which makes them suitable for the supra-segmental approach as well as for the low-level feature frame modeling. Moreover, the networks, as any neural network, are able to handle multidimensional target patterns enabling multitask learning for estimation of a confidence measure and true multidimensional affect prediction. While emotion theory does not tell us explicitly that long look back ranges are necessary, our experience teaches us that in conversation as found in the SEMAINE data, for example, emotion does on the one hand not change too rapidly and on the other hand is dependent on the current context of the discussion (refer to, also Lee et al. [2009]). For example, the four SEMAINE characters are supposed to pull participants into a certain mood (one of the four activation-valence quadrants for each of the characters). As the emotion of some utterances (background noise, poor pronunciation, etc.) might be hard to classify from acoustic features, previous and past utterances might be easier to identify and thus help to clarify the more ambiguous cases in between.

The basic principles of LSTM-RNN are explained in Section 4.1. Next, we introduce the acoustic features in Section 4.2, and describe the proposed method for incremental supra-segmental modeling (Section 4.3) as well as affect modeling on the timescale of low-level feature frames (Section 4.4). In the last part of this section, 4.5, we propose a way of automatically predicting the confidence for the dimensional emotion predictions based on multitask learning of target labels and inter-rater agreement and describe the multitask learning of all five dimensions by one model.

4.1. Long Short-Term Memory Recurrent Neural Networks

As a well-suited technique for online regression of emotion dimensions we consider a specialized Recurrent Neural Network (RNN) architecture called Long Short-Term Memory (LSTM) RNN [Hochreiter and Schmidhuber 1997]. Traditional feed-forward neural networks such as the multilayer perceptron are not suitable for classification of connected time series (especially the low-level feature modeling), as they are static classifiers which classify data frame by frame without considering neighboring frames. In order to use neural networks for classification of connected time series, recurrent networks can be used. There, one or more of the hidden network layers is connected to itself. Thus, the network can learn to model past events by adjusting the weights of the feed-back connection(s).

Analysis of the error flow in traditional recurrent neural nets resulted in the finding that long time lags are inaccessible to existing RNN since the backpropagated error either blows up or decays over time (vanishing gradient problem) [Hochreiter et al. 2001]. This led to the introduction of LSTM-RNNs, which are able to store information in linear memory cells over a longer period of time. An LSTM layer is composed of recurrently connected memory blocks, each of which contains one or more recurrently connected memory cells (refer to Figure 1), along with three multiplicative "gate" units: the input, output, and forget gates.

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. Their effect is to allow the network to store and retrieve information over long periods of time, thereby giving access to long-range context information, which in turn is essential when trying to

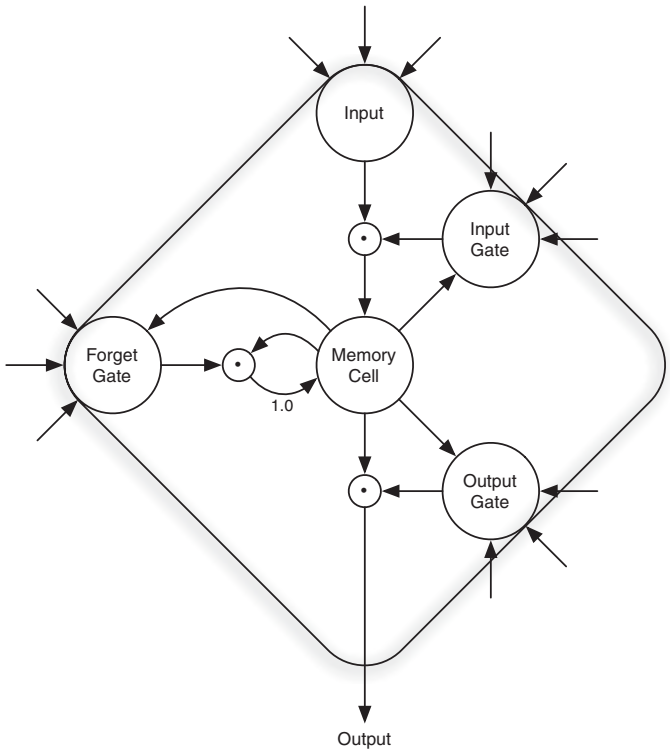

Fig. 1. LSTM memory block consisting of one memory cell: the input, output, and forget gates collect activations from inside and outside the block which control the cell through multiplicative units (depicted as small circles); input, output, and forget gate scale input, output, and internal state respectively; g and h denote activation functions; the recurrent connection of fixed weight 1.0 maintains the internal state.

recognize emotion on a frame level. A more detailed explanation of the LSTM principle can be found in Hochreiter and Schmidhuber [1997].

In LSTM networks, standard feed-forward layers, standard recurrent layers, and LSTM layers can be combined. Thus, a typical network using LSTM memory cells consists of a standard feed-forward input layer with N_i units, where N_i is equal to the number of input features, one or more LSTMs (and optionally standard recurrent) hidden layers consisting of 50–200 memory blocks containing 1–8 LSTM cells each, and one feed-forward output layer with N_o units, where N_o is equal to the number of desired output dimensions or classes.

A further extension of LSTM-RNN is the use of bidirectional networks (see Figure 2), resulting in Bidirectional Long Short-Term Memory Recurrent Neural Networks (BLSTM-RNN) [Schuster and Paliwal 1997]. This method is applied especially for speech recognition tasks [Graves and Schmidhuber 2005; Fernandez et al. 2008], to model anticipatory co-articulation effects. Thereby each hidden layer is duplicated, while one layer processes the inputs forward and the other backward. This results in twice the number of weights in the network, that is, twice the number of parameters to estimate during training. The two hidden layers are connected to the same output layer, which is a standard feed-forward layer and serves the purpose of combining the activations from the forward and the backward hidden layer(s).

One major drawback of this architecture is that the entire input sequence must be available beforehand, which makes this architecture unsuitable for online

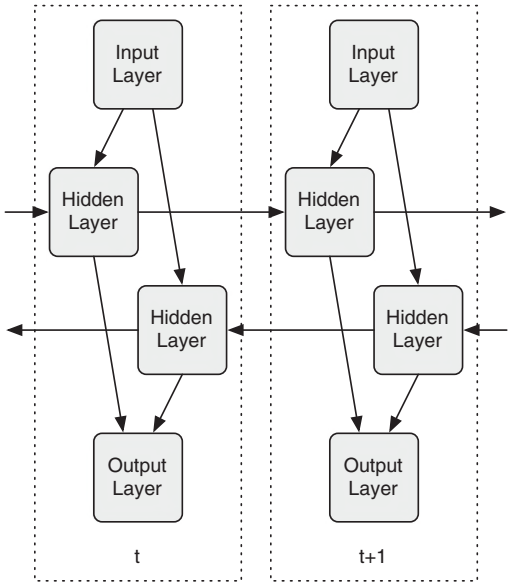

Fig. 2. Bidirectional Recurrent Neural Network.

classification. Therefore, we will not put our primary focus on this architecture, even though we will show by a few exemplary results that this bidirectional network architecture in some cases yields better results than the unidirectional architecture (Section 5). Details on the configurations of the specific networks used for evaluations within this work can be found in Sections 4.3 and 4.4.

LSTM-RNN and BLSTM-RNN can both be trained via standard BackPropagation Through Time (BPTT) [Werbos 1990]. A variant of the standard backpropagation algorithm is resilient Propagation (rProp) [Riedmiller and Braun 1993], where only the sign of the error gradient is considered for network weight updates and *not* the absolute value of the error multiplied by the learn rate. Resilient propagation produces more stable results and thus has outperformed standard backpropagation on many tasks, especially with respect to the number of training iterations required. For affect recognition no study exists which compares the two training algorithms, thus we include a comparison in this article. The error function used as an objective function during training (BPTT and rProp) is the quadratic error regarding the output target (the mean dimension label(s), and/or the inter-rater standard deviation). To avoid overfitting, the best networks are determined by evaluating the correlation coefficients on the development set after each training iteration (epoch), and then choosing the networks which produce the highest correlation coefficient on the development/validation set. The training process is aborted when no improvement over 20 consecutive epochs is observed.

In contrast to the BLSTM-RNN which requires future speech frames, and is therefore more suited for offline processing, the (unidirectional) LSTM-RNN can operate in real time at a real-time factor of 0.085 on an AMD Phenom 64 bit CPU at 2.2 GHz. The asymptotic computational complexity of the LSTM network for recognition of unknown data is $O(n)$ with respect to both number of input samples and feature vector dimensionality. Bidirectional networks have the same complexity (as they are from that point of view nothing else than two unidirectional networks together), but cannot be used in an online system without significant modifications because they require the

Table VI. Feature Set *A* (SEMAINE system, release 3.0 and 3.1)

Low-level descriptors	Functionals
Intensity, Loudness, RMS & LOG energy	Max. and min. value
F_0, prob. of voicing	Range (Max-Min)
MFCC 0–12	Relative position of max. and min. in turn
RASTA-PLP 0–7	Arithmetic mean
log. Mel-Freq. bands 1–14	Linear regression (slope, offset, quadratic/linear error)
95% spectral roll-off point	Standard deviation, Skewness, Kurtosis
Spectral flux, entropy, and variance	% of values $> a \cdot range + min$, $a \in \{0.25, 0.50, 0.75, 0.90\}$
Zero-crossing rate	% of values $< a \cdot range + min$, $a \in \{0.50\}$
	% of rising/falling values

47 low-level descriptors, 20 functionals.

end of a sequence (a 5-minute recording session in our case) before they can process the sequence.

4.2. Acoustic Feature Sets

There is no universally best feature set for affect recognition, and the task and data used is new and quite unexplored by the community. On other databases and tasks many publications exist tackling the issue of feature relevance: Oudeyer [2003] gives a general overview on the topic, Vogt and Andre [2005] compare a broad range of feature sets for acted versus spontaneous emotions, Wu et al. [2010b] investigate different types of acoustic features and find that Mel-Frequency Cepstral Coefficients are among the most relevant feautres, and Batliner et al. [2011] present a quite general, large-scale study contudcted by people from multiple sites aimed at finding relevant features.

As this article is not dealing with finding the best feature set for affect recognition, we use a standard feature set, namely the one we had provided as the official baseline set for the INTERSPEECH 2010 Paralinguistic Challenge. We refer to this as set *B* in the ongoing, and show in comparison to the feature set we have assembled for the SEMAINE demonstrator system (referred to as set *A* in the ongoing), that the choice of the feature set nonetheless is important because set *B* is outperformed by set *A*. This indicates that a more in-depth investigation of the influence of individual features has to be performed in dedicated studies in the future.

Our feature extraction follows the general two-step approach of low-level audio feature extraction followed by subsequent application of functionals to the Low-Level Descriptor (LLD) contours. The low-level audio features are extracted from 25ms windows at a rate of 10ms for all features except the F_0 features, which are extracted from 50ms frames at the same rate. The low-level contours are smoothed with a moving average filter of length 3 frames. LLD contours are either used directly as input to an LSTM-RNN as described in Section 4.4, or functionals are computed from incremental segments as described in Section 4.3. All features have been extracted with our open-source feature extractor openSMILE [Eyben et al. 2010b].

A list of low-level features of set *A* and functionals applied for the supra-segmental modeling are given in Table VI. There are 47 low-level descriptors and 20 functionals applied to all low-level descriptors and their 47 first-order delta coefficients systematically. As two additional features, the number of voiced regions and the segment duration in seconds are considered. In total this results in 1 882 acoustic features in set *A*.

Feature set *B* consists of 1 582 features. The core is the set of 34 low-level descriptors, their 34 delta coefficients multiplied by 21 functionals (1 428 features) as listed in Table VII. For the low-level descriptors raw F_0, jitter, δ jitter, shimmer, the same functionals except for *range* and the 1% percentile (which resemble always 0, i. e., identical to the 99% percentile value, or these LLD when unvoiced segments are present)

Table VII. Feature Set B (INTERSPEECH 2010 paralinguistic Challenge)

Low-level descriptors	Functionals
Loudness	1% and 99% percentile
F_0 envelope, prob. of voicing	Range (1% - 99% percentile)
MFCC 0–14	Relative position of max. and min. in turn
Line Spectral Frequencies 1–8	Arithmetic mean
log. Mel-Freq. bands 1–8	Linear regression (slope, offset, quadratic/linear error)
	Standard deviation, Skewness, Kurtosis
	Quartiles 1–3, inter quartile ranges
	% of values $> 0.75 \cdot range + min$
	% of values $> 0.90 \cdot range + min$

34 low-level descriptors, 21 functionals.

are applied, resulting in an additional $4 \cdot 2 \cdot 19 = 152$ features. The number of voiced regions and the segment duration are added as two extra functionals, resulting in the total number of 1 582 features.

Since training of the LSTM network with a large number of inputs (2–5 k) gives poorer performance in contrast to related work based on Support Vector Machines (see Section 5), we applied a correlation-based feature subset selection (CFS) to the training set to determine five dimension-specific feature selections for each of the two sets. The development and evaluation data is not used in the feature selection process. The CFS algorithm evaluates the worth of each subset of attributes by considering the individual ability of each feature to predict the class or numeric label along with the degree of redundancy between the features. Subsets of features which are highly correlated with the target while having low cross-correlations are preferred. For details please refer to Hall [1998]. In order to compare results of single-task learning to multitask learning, we compute all results on the joint set of selections for all dimensions, as described in Section 4.5.

For feature set A out of 1 882 features, 43 features remain for activation, 46 features for expectation, 23 features for intensity, 34 features for power, and 40 features for valence. For feature set B out of 1 582 features, 38 features remain for activation, 39 features for expectation, 30 features for intensity, 32 features for power, and 28 features for valence. A precise description of these features is difficult to make, as including the full list of selected features for each dimension would be too lengthy. We thus summarize the most frequently occurring low-level descriptors for each dimension in order of their frequency of occurrence (set B only).

—*Activation*: MFCC (16), log. Mel frequency bands (9), LSP frequencies (5), loudness (4), jitter (2).
—*Expectation*: MFCC (18), F_0 (7), LSP frequencies (7), loudness (3), log. Mel frequency bands (2).
—*Intensity*: MFCC (11), loudness (7), LSP frequencies (6) log. Mel frequency bands (5).
—*Power*: MFCC (24), log. Mel frequency bands (3), LSP frequencies (3), F_0 (2).
—*Valence*: MFCC (14), LSP frequencies (7), log. Mel frequency bands (4).

We see that MFCC are always the most frequently selected features (also supported by Wu et al. [2010b]), which we must partially attribute to the fact that MFCC make up a large portion of the original set. Besides MFCC, we can, however, see some variations in selected features among the five dimensions: for activation mostly spectral band energies, format frequencies (related to LSP frequencies), and loudness seem to be important, while for expectation F_0 plays a major role, which seems logical when considering that surprise is the primary emotion category with a low value of expectation; for intensity we see a similar picture as for activation with a slight tendency that loudness seems more important than for activation; for power mostly MFCC-based features

are selected, which might indicate that the subject's dominance is reflected mainly by the way of articulation and less by prosody; valence seems to be characterized best by a mixture of MFCC- and LSP-based features.

The computational complexity of the low-level feature extraction with respect to the number of input frames is always linear ($O(n)$) due to the fact that the descriptors are computed in a single pass on audio signal frames of fixed length. Some might object and say that the required FFT runs at O(n log n), however "n" in this case is the frame size (which is a constant). When assessing the complexity of the feature extraction, "n" refers to the number of frames in the input sequence. Thus, the complexity of the FFT can be seen as a constant factor, and the overall feature extraction algorithm scales linearly. The complexity of the functionals is also linear with respect to the number of segments (5-second windows in this case; see Section 4.3), when the length of the segments is constant. The complexity of the functionals computation with respect to the segment length is linear for all functionals except the percentiles. The algorithm to compute these uses Quicksort, which has a worst-case complexity of $O(n^2)$, and an average complexity of $O(n \log n)$. This means that feature set B can be computed with $O(n \log n)$ with respect to the segment size (linear with respect to all other parameters), and feature set A can be computed with linear complexity with respect to all parameters. The average real-time factor for the complete feature set A on an AMD 64-bit CPU at 2.2GHz is 0.13.

4.3. Incremental Supra-Segmental Modeling

To enable output of emotion predictions at constant time intervals, independent of word- or phrase-level segmentation issues, we decided for a simple, yet powerful incremental segmentation scheme for the supra-segmental approach. As basic unit a speech turn is assumed, which is defined from the point in time where the subject starts talking until the point where the person stops and another person starts talking, or the end of the recording session is encountered. In the SEMAINE database these turns are manually labeled (refer to Section 3), however, in a live system a voice activity detector (optionally in conjunction with a speaker diarization system) can be used instead, which works satisfyingly if tuned properly. The approach subdivides the user's speech turns into overlapping segments with a fixed maximum length. The first segment ranges from $t = 0\,s$ to $t = 1\,s$ with $t = 0\,s$ marking the beginning of the user's speech turn. Should the user's speech turn be smaller than one second the first and only segment ends not at $t = 1\,s$ but at $t = L_T$ with L_T being the length of the turn. For longer turns, the second segment ranges from $t = 0\,s$ to $t = 2\,s$. This is repeated up to the fifth segment from $t = 0\,s$ to $t = 5\,s$. From this point on the segments are kept constant at a length of $5\,s$ and shifted to the right at one-second intervals, that is, the sixth segment ranges from $t = 1\,s$ to $t = 6\,s$. This procedure is now repeated until the end of the turn is reached. The choice of the segment shift of one second is to some extent arbitrary, and the approach can be used to generate outputs at virtually any granularity, to match the needs of the application. However, as the amount of overlap increases, the predictions for the two segments will naturally be more similar.

By applying this method 7 313 turn segments are created from the 1 584 turns in the training set, 1 330 turn segments from the 436 turns in the evaluation set, and 981 turn segments from the 169 turns in the development set. At first the benefit of LSTM-RNN may not seem obvious because the task seems to be a straightforward regression task where the feature vectors for each turn segment can be treated independently. While this is true on the one hand, on the other hand there are temporal dependencies due to the overlap of the segments and the slow changing nature of affect. These dependencies are exploited by LSTM-RNN, which a comparison to standard RNN, feed-forward NN, and SVR shows in Section 5. A complete session, that is, a unit lasting approximately

Table VIII. (B)LSTM-RNN Topologies for Incremental Supra-Segmental Affect Prediction

Topology ID	Bidirectional	Cells in hidden layers
$T1$	no	140, 40
$T1^b$	yes	70, 20
$T2$	no	100, 20
$T2^b$	yes	50,10
$T3$	no	40, 20
$T3^b$	yes	20,10

All networks have two LSTM hidden layers.

5 minutes, where the one user talks to exactly one agent character, is thereby considered as one sequence, that is, an entity which is presented to the network as a connected sequence of inputs. No context is considered across session boundaries.

For each turn segment we evaluate the performance based on acoustic features (set *A*, refer to Section 4.2). The acoustic feature vectors are standardized to have zero mean and unit variance based on statistics collected from the data in the training set.

Six different LSTM-RNN topologies, reflecting differently sized networks, and unidirectional versus bidirectional networks (detailed in Table VIII) are investigated. The selection has been made based on our experience in Eyben et al. [2010a]. However, in contrast to Eyben et al. [2010a] we decided to make the hidden layers in the unidirectional networks twice the size of those in the bidirectional networks, in order to ensure the same number of weights (parameters) in related unidirectional and bidirectional networks. We would like to note that the choice of topologies is by no means meant to be complete and does not substitute a full search over a larger space of network sizes. As this article is not about finding the optimal network topology, we restrict the search to three topologies in order to get a first impression of how large the influence of the network topology is on the performance of multidimensional affect recognition using the SEMAINE data.

During training of the networks Gaussian white noise with standard deviation of 0.3 was added to the input features of the training data. This is a measure to improve the generalization capabilities of neural networks (refer to e.g., Fernandez et al. [2008] and Graves et al. [2005]). It leads to generally longer training times (more epochs), however, avoids overoptimizing on the training data, and thus improves performance on the evaluation and development sets, especially for small databases.

When training neural networks with a gradient descent weight update algorithm, an initial set of weights needs to be chosen, which is unequal to zero. Usually a random initial set of weights is chosen. This makes such kind of training algorithms prone to converge in local minima (of the error target function), depending on the chosen initialization. A common solution to reduce the influence of the initial weights is to train N networks with different initializations for exactly the same problem and take the average of the N output activations. For all experiments reported in this article $N = 5$ runs with random seeds 0–4 for the pseudorandom number generator used to create the initial network weights were performed.

4.4. Low-Level Feature Modeling

Findings in Eyben et al. [2010a] suggest that LSTM-RNN are in principle capable of predicting affect directly from low-level feature descriptors. Although the main focus of this article should be on the proposed incremental supra-segmental approach we compare the supra-segmental approach to the low-level feature-based modeling using the same networks as for the supra-segmental approach. In the case of this low-level feature modeling, the full user speech turns are treated as one sequence. No context

is considered across user turns, due to numeric problems in the training algorithms when handling long sequences, as would be the case when treating a whole session as one sequence. A single topology consisting of two LSTM hidden layers with 140 and 40 units ($T2$), is used. In preliminary experiments larger topologies were investigated. However, due to the rather small dataset (in the light of a such complex learning task) larger networks did not lead to any increase in performance. The network of choice is a unidirectional network, since bidirectional networks would not achieve the goal of low-latency output. Bidirectional networks require the whole input sequence (in our case, one user speech turn) to be present beforehand. We evaluate both types of network though, to asses the performance difference between them.

As the performance of this approach shows that this technology has some potential, but still falls far behind the performance of the supra-segmental approach, we did at this point not investigate it in more detail. Much more work in the future in this direction is required, especially in improving the LSTM-RNN training algorithm and/or investigating other neural network architectures, such as echo state networks, and pooling together or averaging features over a short time segments which have a length that falls between that of the supra-segmental approach (5 seconds) and the low-level frame-based approach (25 ms).

4.5. Multitask Learning

The main novelty of this article is the investigation of multitask learning with LSTM-RNN for prediction of dimensional affect. Previous work of the authors has investigated single-target learning for two affective dimensions only [Eyben et al. 2010a]. Multitask learning is no different from single-task learning, except for the topology of the output layer of the network: For prediction of one continuous dimension an output layer with a single linear summation unit is used; for multitask/multitarget learning the number of linear summation units in the output layer matches the number of targets (2, 5, or 10 in our case). In this article two aspects are investigated: First, instead of training one network for every affect dimension, a single network with five output nodes is used to predict all five dimensions; second, the variance of the four raters' labels (serving as a confidence measure for the dimensional rating) is presented to the network as a second target in addition to the mean of the four raters' labels. Thus—in theory—the network should learn to predict a confidence measure for its output based on the observed input features. Additionally the implicit presentation of the rater agreement information during training might help the network to be able to better predict the dimensional label, as the network could learn to give less weight to more ambiguous training samples, which may improve overall results. In total four configurations of multitask learning are investigated: a single target (rater mean for each dimension), two targets (rater mean and inter-rater standard deviation for each dimension), five targets (rater mean for all five dimensions), and ten targets (rater mean and inter-rater standard deviation for all five dimensions).

A caveat when performing multitask learning is the selection of relevant features. For single-task learning we selected relevant acoustic features for each affect dimension individually with CFS. Instead of adapting the CFS algorithm to be multitarget-capable by averaging of the correlations of each feature with all target labels, we decided to use the joint feature set, that is, the union of the CFS reduced feature sets for all dimensions. This leaves 156 relevant features for feature set A, and 138 for set B. Both numbers are below the sum of the number of features in the respective reduced feature sets (186 and 167), which indicates a small overlap of the reduced feature sets, that is, features that are relevant for more than one dimension.

For feature set *A*, features that are selected for at least 3 dimensions are the skewness of the second MFCC (all four dimensions), the temporal percentage of the regions of

Table IX. Best Results (CC) for the Mean Rater Label Obtained with Given Configurations (topology, feature set, multi-/single-target learning)

Dimension	Feature Set	Topology	Num. targets	CC
A	A	$T3$	2	0.812
E	B	$T2_r^b$ ($T3_r$)	10	0.624 (0.592)
I	A	$T1_r^b$ ($T3$)	10	0.673 (0.656)
P	A	$T1_r^b$ ($T2_r$)	5	0.670 (0.621)
V	B	$T3_r$	1	0.576

Multitarget learning: 2 targets (dimension mean, and inter-rater variance), 5 targets (means of all 5 dimensions), 10 targets (means and inter-rater variances of all 5 dimensions). If the best result is obtained with a bidirectional network, the best unidirectional result is shown in () brackets. ^{b}: Bidirectional network; resilient propagation: r subscript.

rising voicing probability, and the linear error of quadratic regression approximation of the contour of the 13-th log. Mel-frequency band. Including these three, there are 26 features that are selected for at least two dimensions. The most common low-level descriptors among these are the voicing probability, F_0, spectral flux, spectral entropy, MFCC 2 and 6, and the 95% spectral roll-off point.

For feature set B, features that are selected for at least 3 dimensions are the skewness of the third MFCC (all four dimensions), the temporal percentage where the 7-th MFCC is above 75% of its range, and the range of first to second quartile of the 6-th line spectral frequency, as well as the second quartile of the 7-th line spectral frequency. Including these three, there are 24 features that are selected for at least two dimensions. The most common low-level descriptors among these are the loudness, 0-th and 6-th line spectral frequency, and MFCC 10, as well as the voicing probability.

Please note, as previously mentioned, we conduct all experiments for single-task and multitask learning on the joint feature set to ensure comparability of results and eliminate the influence of different feature sets.

5. EVALUATION AND EXPERIMENTAL RESULTS

Extensive results of the evaluations of the incremental supra-segmental modeling are presented and discussed in this section. The supra-segmental approach is compared to the low-level feature modeling approach. As measure of evaluation we report the Correlation Coefficient (CC) between the automatic predictions and the ground-truth labels (mean label of raters), as suggested in Schuller et al. [2010b] and applied in Eyben et al. [2011, 2010a].

A large number of results have been obtained for all the runs evaluating the performance for the five affective dimensions with 12 (B)LSTM topologies and 2 RNN as well as 2 NN topologies, two acoustic feature sets, and 3 multitask learning setups versus single-task learning. In total results for 640 individual runs were computed. Each individual result was produced by training 5 networks with different initial weights on the same data, and averaging the output activations over those 5 networks. This is done to lessen the influence of the training converging into local minima and make results more stable. To give a meaningful and informative summary of the individual results we report averaged results in terms of average correlation coefficients over various configurations. In particular average correlation coefficients for each topology (Table XI), each feature set (Table XII), and each multi-/single-task learning variant (Table XIII) are given. The overall best individual results are shown in Table IX. A comparison to related approaches is given in Table X. The related methods include Support Vector Regression (SVR), standard feed-forward Neural Networks (NN), and standard Recurrent Neural Networks (RNN), both having the same size as the $T1$

Table X. Comparison of LSTM-RNN to Support Vector Regression (SVR), Feed-Forward Neutral Networks (NN), and Standard Recurrent Neural Networks (RNN)

Dimension	LSTM (CC)	RNN (CC)	NN (CC)	SVR (CC)
A	**0.757**	0.725	0.709	0.653
E	**0.549**	0.302	−0.029	0.190
I	**0.579**	0.518	0.461	0.503
P	**0.520**	0.511	0.361	0.367
V	**0.454**	0.172	0.035	−0.085

Mean rater label CC. Topology $T2_r$, single target (mean of each affective dimension), feature set *B*. LSTM-RNN, RNN, and NN have the same number of hidden units in 2 hidden layers and are all trained with resilient propagation (topology $T2_r$). Best result marked in boldface font.

LSTM-RNN: two hidden layers with 140 and 40 summation units (sigmoid transfer function), respectively.

Networks trained with resilient propagation are marked with the r subscript (i. e., $T2_r$ is a topology $T2$ network trained with resilient propagation). All other networks were trained with backpropagation through time. We see a clear trend which is well known throughout the literature (e.g., Grimm et al. [2007a] and Eyben et al. [2010a]), that the *valence* (V) dimension performs worst, while *activation* (A) performs best. The new dimensions *expectation* (E), *intensity* (I), and *power* (P) perform fairly well, in terms of CC. We decided to base our analysis on correlation coefficient as the evaluation metric only, as the only other commonly used metrics Mean-Squared Error (MSE) and the Mean Linear Error (MLE) are disturbed by scaling and bias in the neural network outputs. This is based on our experience from Eyben et al. [2010a] where we showed that the outputs of the neural networks are often correlated to the targets, but show a bias and/or scaling. We thus prefer the correlation coefficient as measure of choice.

The best results from the 640 runs and the respective configurations are shown in Table IX. For all dimensions except activation the best result is obtained with a bidirectional network. Best unidirectional networks are approximately 0.02 to 0.07 behind the B-LSTM. Concerning topologies there is no clear trend apparent from this table. Resilient propagation as training algorithm leads to better networks in slightly more cases. Feature set *A* wins in most cases, except for expectation and valence. Comparing the best results obtained with (B)LSTM networks in Table IX with the average correlation between the human raters on the test set in Table IV, we can see that the automatic system actually outperforms the human performance for all dimensions except valence. This is mostly in line with the findings reported in Eyben et al. [2011]. Valence has the best agreement among human raters but is most difficult to predict for the automatic system relying on acoustic cues only. A lot of valence information is carried by the linguistic content and the context of an utterance. Given the low human agreement of the dimensions *expectation* and *power*, the results of automatic recognition seem very good, which is surprising, but seems to show that the average of the human ratings does provide a ground truth which is correlated to some acoustic properties and thus more valid than one would assume from the high rater ambiguity.

A comparison of LSTM-RNN to other related neural networks is shown in Table X. The RNN and NN have the same number of hidden units in two hidden layers as the LSTM-RNN. Support Vector Regression has been trained with the Sequential Minimal Optimization (SMO) algorithm using the WEKA toolkit [Witten and Frank 2005]. Thereby a linear kernel function was used and the complexity parameter c was chosen as 1.0. In all five cases the LSTM outperforms the other methods. For activation,

Table XI. Average Mean Rater Label CC per Topology

Dimension	CC-1	CC-10
$T1$	0.513	0.530
$T1^b$	0.543	0.552
$T1_r$	0.551	0.585
$T1_r^b$	**0.583**	**0.590**
$T2$	0.506	0.501
$T2^b$	0.512	0.523
$T2_r$	0.524	0.544
$T2_r^b$	0.566	0.582
$T3$	0.526	0.481
$T3^b$	0.499	0.500
$T3_r$	0.577	0.564
$T3_r^b$	0.583	0.563
$T2^{nn}$	0.368	0.430
$T2^{rnn}$	0.475	0.458
$T2_r^{nn}$	0.368	0.387
$T2_r^{rnn}$	0.462	0.433

CC averaged over both feature sets and all five affective dimensions. **CC-1**: CC averaged over single-target runs (dimension mean as target); **CC-10**: CC averaged over multitarget runs (mean and inter-rater variance of all five dimensions as targets). Bottom part: feed-forward and standard recurrent neural network topologies trained with backpropagation through time and resilient propagation (r subscript). ^{b}: Bidirectional network.

intensity, and power the performance of LSTM and standard RNN can be seen as identical, indicating that long-term context is not of great importance, whereas for expectation and valence the LSTM significantly outperform the RNN and especially the NN, which is an indicator for the importance of long-term context here. Except for intensity and expectation the performance of SVR is behind NN. The performance of NN is always behind that of RNN. For intensity RNN, NN, and SVR all yield a CC of 0.5, which is roughly 0.07 behind the CC obtained with LSTM.

In order to find out which topology performs best, we averaged results over all dimensions and feature sets. The results can be seen in Table XI. No clearly best performing topology can be identified, as all topologies achieve an average CC close to 0.5. The topology $T1_r^b$ is marginally the best. Since $T1$ is the largest network size, this suggests that bigger networks are to be preferred, but at the same time the significantly smaller networks ($T3$) do not perform too badly, especially when trained with resilient propagation. A very clear trend among all topologies is visible indicating that networks trained with resilient propagation perform better than those trained with backpropagation through time. Bidirectional networks (even though they have exactly the same amount of parameters as the corresponding unidirectional networks) perform better among all three topologies. Table XI confirms what was already discussed earlier, namely that the NN and RNN networks perform worse than the LSTM networks. This again shows that LSTM brings a benefit for the task of dimensional affect recognition. A final observation we can make from Table XI is that the larger topologies ($T1$ and $T2$) perform better for multitask learning and the smaller topology ($T3$) performs better for single-task learning, which is to be expected, as generally speaking a model for multitask learning must be able to hold more parameters.

Overall, the differences between the topologies are not very large. From this finding we can assume that a further, more fine-grained investigation of further topologies is not necessary and/or will not yield to any significant improvement over the current results. For future studies the most interesting experiment in this respect is to further

Table XII. Average Result for Each Feature Set (CC), Averaged over All LSTM Topologies and All Five Affective Dimensions

Feature Set	**A**	**B**
CC	0.535	0.519

Table XIII. Average CC Evaluated for the Mean Rater Label (over all dimensions, topologies, and feature sets) for Various Number of Targets

	CC
CC-1	0.529 (0.487)
CC-2	0.533 (0.447)
CC-5	0.529
CC-10	0.543

CC-1: CC averaged over all single-target runs (dimension mean as target); **CC-2**: CC averaged over all single-target runs (dimension and inter-rater variance as target); **CC-5**: CC averaged over multitarget runs (mean of all five dimensions as targets); **CC-10**: CC averaged over multitarget runs (mean and inter-rater variance of all five dimensions as targets). Results shown in () brackets are results achieved with an individual feature selection per dimension, the other results are those obtained on the same feature set as the multitask results.

reduce the network size to find the point where performance significantly drops, and thus determine a minimal network size for the task.

The fact that most best results were obtained with feature set *A* (Table IX) can be confirmed in Table XII, where the averaged result per feature set is shown. These results were obtained by averaging all individual results (single and multitask, all topologies) for each feature set. Although the difference between the two feature sets is not great, feature set *A* yields a by 0.01 higher average CC. However, we have to interpret this fact with care, as feature set *A* is the larger set of both (156 features versus 138 features), and it is a general trend that more features perform better than fewer up to the point where data sparseness becomes an issue. We can confirm this trend when looking explicitly at results of single-task learning obtained on the per-dimension feature selections (refer to Section 4.2) as opposed to the union of per-dimension feature sets (Table XIII). The smaller feature set, but the one that should be optimized for each dimension, in theory, shows a 0.04 lower CC for single-task learning, and a 0.08 lower CC for 2-task learning (target and inter-rater variance).

A similar difference in performance can be seen for single-task versus multitask learning. Table XIII shows the results for the four cases of: (a) single-task learning of each dimension individually, (b) two-task learning of each dimension and corresponding inter-rater standard deviation as a confidence measure, (c) multitask learning of all 5 dimensions, and (d) multitask learning of all 5 dimensions and corresponding inter-rater standard deviations. Including the inter-rater standard deviation improves the results marginally (CC 0.01), while multitask learning of all five dimensions does not seem to have a great effect on average (individual cases differ). The winning way of modeling 5 continuous affective dimensions is by modeling all 5 dimensions and the corresponding inter-rater standard deviations in a single network. A possible explanation for the fact that inter-rater standard deviation improves results more for multidimension learning could be that the uncertainty information added by all dimensions helps the network. This hypothesis is further underscored by the fact that the best correlation coefficients for prediction of the inter-rater standard deviations

Table XIV. Best Result Obtained for Prediction of Inter-Rater Standard Deviation

Dimension	Feature Set	Topology	CC
A	A	$T3_r^b$	0.241
E	B	$T1$	0.306
I	B	$T3_r^b$	0.237
P	A	$T3^b$	0.412
V	A	$T2^b$	0.125

All results obtained with 2-target learning (mean and standard deviation of raters for each dimension). ^{b}: Bidirectional network; resilient propagation: r subscript.

Table XV. Correlation of Loudness and F_0 with the Five Dimensional Labels (mean of raters)

[CC]	max. loudness	mean. loudness	stddev. loudness
A	0.65	0.63	0.60
E	0.06	−0.03	0.10
I	0.57	0.49	0.52
P	−0.00	0.11	−0.03
V	0.16	−0.12	−0.12
	max. F_0	mean. F_0	stddev. F_0
A	0.39	0.50	0.22
E	0.09	0.18	0.20
I	0.33	0.49	0.26
P	−0.08	−0.15	−0.18
V	−0.04	−0.09	0.01

Statistics mean, maximum, and standard deviation computed over the incremental (5-second) segments.

have been obtained with single-dimension learning and not multidimension learning. These results are shown in Table XIV. The uncertainty of intensity and power seem to be most easily predictable. For the other dimensions the prediction of the uncertainty is fairly poor.

Thus, concluding, to use the predicted inter-rater standard deviation as an actual confidence measure for all dimensions more work is required to optimize this prediction. However, with the multitask approach presented herein, it is beneficial to include the inter-rater standard deviation in order to improve the prediction of the primary target, the mean label for each dimension.

In order to justify the use and feasibility of any sort of classifier/regressor on the SEMAINE data for dimensional affect recognition, we have computed the correlation coefficients between three functionals of the low-level acoustic features loudness and F_0 and the mean labels for the five dimensions on the evaluation set. The three functionals are mean, maximum, and standard deviation within a segment. The segments are the same (overlapping) segments as used in our proposed incremental supra-segmental approach. Table XV shows these correlation coefficients. The result is very interesting, as for some features and the dimensions activation and intensity very high CC are obtained, while for the other three dimensions no significant correlation can be reported. The maximum loudness per segment yields a correlation coefficient of 0.65 with the activation dimension. This is above the average human rater agreement (0.57) but below the best result obtained with LSTM (0.81), thus justifying the use of LSTM. Further, the finding is in line with the result of the feature selection (Section 4.2), which revealed loudness-related features highly relevant for activation and intensity. For loudness the maximum loudness per segment seems to be better correlated to activation and intensity, while for F_0, the mean F_0 per segment shows a stronger correlation.

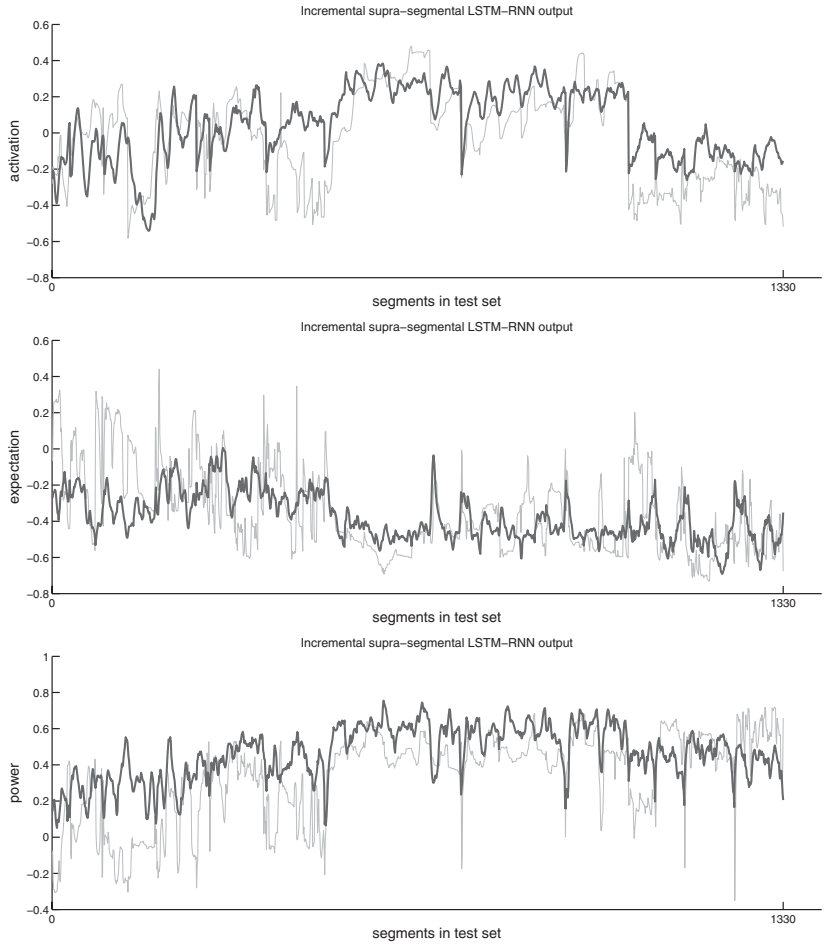

Fig. 3. Plots of predictions for the configurations that gave the best results for activation (top), expectation (middle), and power (bottom) (Table IX). The thick (blue) line is the result of the automatic prediction, the thin (green) line is the ground truth (mean of raters). On the x-axis all segments in the evaluation set are arranged in chronological order.

A small correlation between the expectation dimension and the F_0 standard deviation (0.2) is observed.

Besides looking at the correlation coefficients, the best way to judge the actual performance of the networks and to analyse what is actually happening to the outputs is to take a look at the plots shown Figure 3. They show the actual network outputs obtained with the networks that gave the best results as shown in Table IX. All user turns are concatenated and the gaps created by operator turns in between are not shown in order to keep the plot clean and easy to read.

The results for the low-level feature-based modeling of affective dimensions are given in Table XVI. The correlations obtained with this approach are very low compared to the supra-segmental approach and the simple feature to label correlation presented in Table XV. Thus, we conclude that at present the supra-segmental modeling should be the preference and the low-level modeling needs more investigation and improvements. No clear trend showing a best configuration can be seen from Table XVI, and the

Table XVI. LLD-Level Modeling of Mean Rater Label for 3 Dimensions

Configuration	**A**	**E**	**V**
1-dim learning (LSTM-rp)	0.082	0.355	−0.006
1-dim learning (BLSTM-rp)	0.123	0.323	0.003
1-dim learning (BLSTM-bptt)	0.271	0.279	0.090
5-dim learning (LSTM-rp)	0.560	0.110	0.116
5-dim learning (BLSTM-rp)	0.469	0.056	0.295
5-dim learning (BLSTM-bptt)	0.378	0.056	0.296

Results obtained with LSTM, topology $T1$, resilient propagation (rp) or backpropagation through time (bptt) for training. Dimensions A(ctivation), E(xpectation), V(alence). Correlation coefficient averaged over 5 network trainings with different initial weights (same procedure as for the supra-segmental results). Uni- and bidirectional LSTM (LSTM/BLSTM). Single-task (1-dim) learning compared to multitask learning (5-dim).

performance for activation, for example, shows very high variance (CC 0.082 to 0.560), which might be an indication of instabilities of the training algorithms in the case of this complex task. Therefore, besides optimizing network topology in future work, using other training algorithms, such as the Extended Kalman Filter (EKF) training [Pérez-Ortiz et al. 2003] might seem promising in this respect.

6. CONCLUSIONS

We have presented a novel incremental segmentation scheme for supra-segmental modeling of multidimensional affect from acoustic cues, which is suitable for low-latency, spontaneous, and naturalistic affect estimation in realistic environments. The approach uses Long Short-Term Memory Recurrent Neural Networks for multitask modeling. This article is the first to investigate the joint learning of affective dimensions. Various network topologies were compared, including bidirectional and unidirectional networks. Due to the incremental output during a user's speech turn the approach is suitable for use in real virtual agents and robots. The SEMAINE database is used for experiments, which contains spontaneous and natural interactions of humans with four emotionally stereotypical Wizard-of-Oz characters. Five affective dimensions are annotated in this database and correlation coefficients of up to 0.81 for activation, 0.62 for expectation, 0.67 for intensity, 0.67 for power, and 0.58 for valence are reported. Thereby LSTM outperformed standard recurrent neural networks, feed-forward neural networks, and Support Vector Regression by 0.1 average correlation coefficient. No clear tendency towards an optimal network topology was found, however, standard backpropagation trained networks were found to yield inferior correlation coefficient but produce outputs more in the proper range than networks trained by resilient propagation, which in turn yield a higher correlation coefficient. Considering that resilient propagation only uses the sign of the error function for weight updates this result is explicable.

Further, we have suggested a novel approach for estimating confidences of continuous dimensional affect predictions by multitask learning of the mean of the raters along with the standard deviation of the raters. When learning the standard deviations and the means of all five dimensions with one network, a benefit can be shown which is attributed to the additional labels of inter-rater standard deviation. The prediction of the confidences by themselves is feasible for some configurations, but requires far more tuning and a more in-depth study in order to advance the method to a state where reliable confidences can be obtained.

Concluding, we can say that realistic, natural affect recognition is getting towards a state where it can be used in real-world intelligent affective systems. Future work

shall encompass the investigation of alternate, and more stable training algorithms for the LSTM networks, such as Extended Kalman Filter training. Moreover, the fusion of acoustic and linguistic features, which was proven successful, especially for valence, in Eyben et al. [2010a], shall be combined with the herein presented approach of multitask learning. Next, the fusion of multitask learning of acoustic and linguistic cues together with visual features will be investigated, leading to audiovisual affect recognition.

REFERENCES

BATLINER, A., SEPPI, D., STEIDL, S., AND SCHULLER, B. 2010. Segmenting into adequate units for automatic recognition of emotion-related episodes: A speech-based approach. *Advan. Hum. Comput. Interact.* article ID 782802.

BATLINER, A., STEIDL, S., SCHULLER, B., SEPPI, D., VOGT, T., WAGNER, J., DEVILLERS, L., VIDRASCU, L., KESSOUS, V. A. L., AND AMIR, N. 2011. Whodunnit—Searching for the most important feature types signalling emotion-related user states in speech. *Comput. Speech Lang. 25,* 1.

BURKHARDT, F., PAESCHKE, A., ROLFES, M., SENDLMEIER, W., AND WEISS, B. 2005. A database of german emotional speech. In *Proceedings of the Interspeech Conference*. 1517–1520.

BUSSO, C., LEE, S., AND NARAYANAN, S. S. 2007. Using neutral speech models for emotional speech analysis. In *Proceedings of the Interspeech Conference*. 2225–2228.

CARIDAKIS, G., MALATESTA, L., KESSOUS, L., AMIR, N., RAOUZAIOU, A., AND KARPOUZIS, K. 2006. Modeling naturalistic affective states via facial and vocal expressions recognition. In *Proceedings of the ACM International Conference on Multimodal Interfaces*. 146–154.

CHANG, C.-C. AND LIN, C.-J. 2001. *LibSVM: A Library for Support Vector Machines*. `http://www.csie.ntu.edu.tw/cjlin/libsvm`.

COHEN, J., COHEN, P., WEST, S. G., AND AIKEN, L. S. 2003. *Applied Multiple Regression / Correlation Analysis for the Behavioral Sciences*, 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

COWIE, R., DOUGLAS-COWIE, E., SAVVIDOU, S., MCMAHON, E., SAWEY, M., AND SCHRÖDER, M. 2000. Feeltrace: an instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion*. 19–24.

DEVILLERS, L., VIDRASCU, L., AND LAMEL, L. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Netw. 18,* 4, 407–422.

DOUGLAS-COWIE, E., COWIE, R., SNEDDON, I., COX, C., LOWRY, O., MCRORIE, M., MARTIN, J. C., DEVILLERS, L., ABRILIAN, S., BATLINER, A., AMIR, N., AND KARPOUZIS, K. 2007a. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Affective Computing and Intelligent Interaction*. Springer, 488–500.

DOUGLAS-COWIE, E., COWIE, R., SNEDDON, I., COX, C., O., L., MCRORIE, M., MARTIN, J., DEVILLERS, L., ABRILIAN, S., BATLINER, A., AMIR, N., AND KARPOUZIS, K. 2007b. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Lecture Notes in Computer Science*,Springer, vol. 4738, 488–501.

EKMAN, P. AND FRIESEN, W. V. 1975. *Unmasking the Face: A Guide to Recognizing Emotions from Facial Expressions*. Prentice Hall, Englewood Cliffs, NJ.

EYBEN, F., WÖLLMER, M., GRAVES, A., SCHULLER, B., DOUGLAS-COWIE, E., AND COWIE, R. 2010a. On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues. *J. Multimodal User Interfaces 3,* 1-2, 7–19.

EYBEN, F., WÖLLMER, M., AND SCHULLER, B. 2010b. openSMILE—The Munich versatile and fast open-source audio feature extractor. In *Proceedings of ACM Multimedia Conference*. 1459–1462.

EYBEN, F., WÖLLMER, M., VALSTER, M., GUNES, H., SCHULLER, B., AND PANTIC, M. 2011. String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In *Proceedings FG Conference* (to appear).

FERNANDEZ, S., GRAVES, A., AND SCHMIDHUBER, J. 2008. Phoneme recognition in timit with blstm-ctc. Tech. rep., IDSIA.

FONTAINE, J. R. J., SCHERER, K. R., ROESCH, E. B., AND ELLSWORTH, P. C. 2007. The world of emotions is not two-dimensional. *Psychol. Sci. 18,* 2, 1050–1057.

FRAGOPANAGOS, N. AND TAYLOR, J. G. 2005. Emotion recognition in human-computer interaction. *Neural Netw. 18,* 4, 389–405.

GLOWINSKI, D., CAMURRI, A., VOLPE, G., DAEL, N., AND SCHERER, K. 2008. Technique for automatic emotion recognition by body gesture analysis. In *Proceedings of Computer Vision and Pattern Recognition Workshops*. 1–6.

GRAVES, A., FERNANDEZ, S., AND SCHMIDHUBER, J. 2005. Bidirectional lstm networks for improved phoneme classification and recognition. In *Proceedings of ICANN*. Vol. 18. 602–610.

GRAVES, A. AND SCHMIDHUBER, J. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Netw. 18,* 5-6, 602–610.

GRIMM, M., KROSCHEL, K., AND NARAYANAN, S. 2007a. Support vector regression for automatic recognition of spontaneous emotions in speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Vol. 4. IEEE.

GRIMM, M., MOWER, E., KROSCHEL, K., AND NARAYANAN, S. 2007b. Primitives based estimation and evaluation of emotions in speech. *Speech Comm. 49*, 787–800.

GUNES, H. AND PANTIC, M. 2010a. Automatic, dimensional and continuous emotion recognition. *Int. J. Synth. Emot. 1,* 1, 68–99.

GUNES, H. AND PANTIC, M. 2010b. Automatic measurement of affect in dimensional and continuous spaces: Why, what, and how? In *Proceedings of the Conference on Measuring Behavior*. 122–126.

GUNES, H. AND PANTIC, M. 2010c. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In *Proceedings of International Conference on Intelligent Virtual Agents*. 371–377.

HALL, M. A. 1998. Correlation-based feature subset selection for machine learning. Ph.D. thesis, University of Waikato, Hamilton, New Zealand.

HOCHREITER, S., BENGIO, Y., FRASCONI, P., AND SCHMIDHUBER, J. 2001. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In *A Field Guide to Dynamical Recurrent Neural Networks*, S. C. Kremer and J. F. Kolen, Eds., IEEE Press.

HOCHREITER, S. AND SCHMIDHUBER, J. 1997. Long short-term memory. *Neural Comput. 9*, 8, 1735–1780.

IOANNOU, S., RAOUZAIOU, A., TZOUVARAS, V., MAILIS, T., KARPOUZIS, K., AND KOLLIAS, S. 2005. Emotion recognition through facial expression analysis based on a neurofuzzy method. *J. Neural Netw. 18*, 423–435.

LEE, C., BUSSO, C., LEE, S., AND NARAYANAN, S. 2009. Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. In *Proceedings of the Interspeech Conference*. 1983–1986.

LICHTENAUER, J., SHEN, J., VALSTAR, M., AND PANTIC, M. 2010. Cost-Effective solution to synchronised audio-visual data capture using multiple sensors. *J. Vis. Comm. Image Represent.*, 1–39.

MCKEOWN, G., VALSTAR, M. F., PANTIC, M., AND COWIE, R. 2010. The semaine corpus of emotionally coloured character interactions. In *Proceedings of the ICME Conference*. IEEE, 1–6.

METALLINOU, A., KATSAMANIS, A., WANG, Y., AND NARAYANAN, S. S. 2011. Tracking changes in continuous emotion states using body language and prosodic cues. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2288–2291.

MOWER, E., MATARIC, M. J., AND NARAYANAN, S. S. 2011. A framework for automatic human emotion classification using emotional profiles. *IEEE Trans. Audio, Speech Lang. Process. 19,* 5, 1057–1070.

MOWER, E. AND NARAYANAN, S. S. 2011. A hierarchical static-dynamic framework for emotion classification. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2372–2375.

NICOLAOU, M., GUNES, H., AND PANTIC, M. 2010. Audio-visual classification and fusion of spontaneous affective data in likelihood space. In *Proceedings of IEEE International Conference on Pattern Recognition*. 3695–3699.

OUDEYER, P. Y. 2003. The production and recognition of emotions in speech: Features and algorithms. *Int. J. Hum.-Comput. Studies 59*, 157–183.

PÉREZ-ORTIZ, J. A., GERS, F. A., ECK, D., AND SCHMIDHUBER, J. 2003. Kalman filters improve lstm network performance in problems unsolvable by traditional recurrent nets. *Neural Netw. 16,* 2, 241–250.

PETERS, C. AND O'SULLIVAN, C. 2002. Synthetic vision and memory for autonomous virtual humans. *Comput. Graph. Forum 21,* 4, 743–753.

RIEDMILLER, M. AND BRAUN, H. 1993. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *Proceedings of the IEEE International Conference on Neural Networks*. 586–591.

SCHMIDT, E. M. AND KIM, Y. E. 2010. Prediction of time-varying musical mood distributions from audio. In *Proceedings of the International Society for Music Information Retrieval Conference*.

SCHULLER, B., BATLINER, A., STEIDL, S., AND SEPPI, D. 2010a. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Comm. (Special Issue on Sensing Emotion and Affect Facing Realism in Speech Processing)* (to appear).

SCHULLER, B., MÜLLER, R., EYBEN, F., GAST, J., HÖRNLER, B., WÖLLMER, M., RIGOLL, G., HÖTHKER, A., AND KONOSU, H. 2009a. Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. *Image Vis. Comput. J. 27,* 12, 1760–1774.

SCHULLER, B., REITER, S., AND RIGOLL, G. 2006. Evolutionary feature generation in speech emotion recognition. In *Proceedings of the International Conference on Multimedia and Expo (ICME)*. 5–8.

SCHULLER, B. AND RIGOLL, G. 2006. Timing levels in segment-based speech emotion recognition. In *Proceedings of the Interspeech Conference*. 1818–1821.

SCHULLER, B., RIGOLL, G., AND LANG, M. 2003. Hidden markov model-based speech emotion recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. Vol. II. IEEE, 1–4.

SCHULLER, B., SEPPI, D., BATLINER, A., MAIER, A., AND STEIDL, S. 2007a. Towards more reality in the recognition of emotional speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. Vol. IV. IEEE, 941–944.

SCHULLER, B., STEIDL, S., AND BATLINER, A. 2009b. The INTERSPEECH 2009 emotion challenge. In *Proceedings of the Interspeech Conference*. 312–315.

SCHULLER, B., STEIDL, S., BATLINER, A., BURKHARDT, F., DEVILLERS, L., MÜLLER, C., AND NARAYANAN, S. 2010b. The interspeech 2010 paralinguistic challenge. In *Proceedings of the Interspeech Conference*. 2794–2797.

SCHULLER, B., STEIDL, S., BATLINER, A., SCHIEL, F., AND KRAJEWSKI, J. 2011. The interspeech 2011 speaker state challenge. In *Proceedings of the Interspeech Conference*.

SCHULLER, B., VLASENKO, B., EYBEN, F., RIGOLL, G., AND WENDEMUTH, A. 2009c. Acoustic emotion recognition: A benchmark comparison of performances. In *Proceedings of the ASRU Conference*. IEEE.

SCHULLER, B., VLASENKO, B., EYBEN, F., WÖLLMER, M., STUHLSATZ, A., WENDEMUTH, A., AND RIGOLL, G. 2010c. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Trans. Affective Comput. 1,* 2.

SCHULLER, B., VLASENKO, B., MINGUEZ, R., RIGOLL, G., AND WENDEMUTH, A. 2007b. Comparing one and two-stage acoustic modeling in the recognition of emotion in speech. In *Proceedings of the Automatic Speech Recognition and Understanding Workshop*. 596–600.

SCHULLER, B., WIMMER, M., MÖSENLECHNER, L., KERN, C., ARSIC, D., AND RIGOLL, G. 2008. Brute-Forcing hierarchical functionals for paralinguistics: A waste of feature space? In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*. 4501–4504.

SCHULLER, B., ZACCARELLI, R., ROLLET, N., AND DEVILLERS, L. 2010d. Cinemo a french spoken language resource for complex emotions: Facts and baselines. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*.

SCHUSTER, M. AND PALIWAL, K. K. 1997. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process. 45*, 2673–2681.

STEIDL, S. 2009. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos, Berlin.

STEIDL, S., SCHULLER, B., BATLINER, A., AND SEPPI, D. 2009. The hinterland of emotions: Facing the open-microphone challenge. In *Proceedings of the 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction (ACII)*. Vol. I. IEEE, 690–697.

STREIT, M., BATLINER, A., AND PORTELE, T. 2006. Emotions analysis and emotion-handling subdialogues. In *SmartKom: Foundations of Multimodal Dialogue Systems*, W. Wahlster, Ed., Springer, 317–332.

VLASENKO, B., SCHULLER, B., WENDEMUTH, A., AND RIGOLL, G. 2007. Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing. In *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction (ACII)*, Lecture Notes in Computer Science, vol. 4738. Springer, 139–147.

VLASENKO, B. AND WENDEMUTH, A. 2007. Tuning hidden markov model for speech emotion recognition. In *Proceedings of DAGA 33rd German Annual Conference on Acoustics*.

VOGT, T. AND ANDRE, E. 2005. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *Proceedings of the ICME Conference*. 474–477.

WERBOS, P. 1990. Backpropagation through time: What it does and how to do it. *Proc. IEEE 78*, 1550–1560.

WITTEN, I. H. AND FRANK, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. Morgan Kaufmann, San Francisco.

WÖLLMER, M., EYBEN, F., REITER, S., SCHULLER, B., COX, C., DOUGLAS-COWIE, E., AND COWIE, R. 2008. Abandoning emotion classes - Towards continuous emotion recognition with modelling of long-range dependencies. In *Proceedings of the Interspeech Conference*. 597–600.

WÖLLMER, M., EYBEN, F., SCHULLER, B., DOUGLAS-COWIE, E., AND COWIE, R. 2009. Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks. In *Proceedings of Interspeech Conference*. 1595–1598.

WÖLLMER, M., METALLINOU, A., EYBEN, F., SCHULLER, B., AND NARAYANAN, S. 2010a. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proceedings of Interspeech Conference*. 2362–2365.

WÖLLMER, M., SCHULLER, B., EYBEN, F., AND RIGOLL, G. 2010b. Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening. *IEEE J. Select. Topics Signal Process. 4,* 5, 867–881.

WU, D., PARSONS, T., MOWER, E., AND NARAYANAN, S. S. 2010a. Speech emotion estimation in 3d space. In *Proceedings of the ICME Conference*. 737–742.

WU, D., PARSONS, T., AND NARAYANAN, S. S. 2010b. Acoustic feature analysis in speech emotion primitives estimation. In *Proceedings of the Interspeech Conference*. 785–788.

YEE, P. V. AND HAYKIN, S. 2001. *Regularized Radial Basis Function Networks: Theory and Applications*. John Wiley.

ZENG, Z., PANTIC, M., ROISMAN, G. I., AND HUANG, T. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell. 31,* 1, 39–58.