On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues

Florian Eyben · Martin Wöllmer · Alex Graves · Björn Schuller · Ellen Douglas-Cowie · Roddy Cowie

Abstract For many applications of emotion recognition, such as virtual agents, the system must select responses while the user is speaking. This requires reliable on-line recognition of the user's affect. However most emotion recognition systems are based on turnwise processing. We present a novel approach to on-line emotion recognition from speech using Long Short-Term Memory Recurrent Neural Networks. Emotion is recognised frame-wise in a two-dimensional valence-activation continuum. In contrast to current state-of-the-art approaches, recognition is performed on low-level signal frames, similar to those used for speech recognition. No statistical functionals are applied to low-level feature contours. Framing at a higher level is therefore unnecessary and regression outputs can be produced in real-time for every low-level input frame. We also investigate the benefits of including linguistic features on the signal frame level obtained by a keyword spotter.

F. Eyben M. Wöllmer · B. Schuller
Institute for Human-Machine Communication,
Technische Universität München, Theresienstrasse 90,
80333 Munich, Germany
e-mail: eyben@tum.de

A. Graves
Institute for Computer Science VI, Technische Universität
München, Boltzmannstrasse 3, 85748 Munich, Germany

E. Douglas-Cowie · R. Cowie
School of Psychology, Queen's University, Belfast BT7 1NN, UK

1 Introduction

Emotionally sensitive virtual agents need to be able to estimate a user's emotion in real-time. However most Automatic Emotion Recognition (AER) systems are designed for supra-segmental units of speech [50], such as complete sentences or fragments of sentences [30, 44]. This means that no output is generated by the system until the user pauses for a long time, and that changes in emotion during a segment are ignored. Furthermore the fragments—in most cases—are pre-segmented and all results obtained have the precondition of perfect segmentation. However, in a real-world, on-line AER system, the segmentation is not known and methods such as Voice Activity Detection (VAD) and energy thresholding must be implemented. These, however, will not achieve perfect segmentation on, e.g. the sentence level. All those aspects highlight that it is extremely important to close or at least narrow the gap between the human ability to permanently observe, question, update, and refine the estimation of the current affect of a conversational partner, and the simple evaluation of features at the end of a speech turn as it is applied for many virtual agents (e.g. [40])—even though this does not mean that human perception is continuous at all points.

In this article, we introduce a novel approach to fully continuous emotion recognition in a 3-D valence-activation-time continuum. The novelty of our technique is that it explicitly models *time* as a third dimension—similar to automatic speech recognition systems.

The motivation for time-continuous emotion recognition is clearly given by applications where emotion must be evaluated incrementally in real-time, e.g. in-car driver monitoring for safety enhancement, or virtual agents as in our case: in the SEMAINE system [27] customised and immediate feedback based on the emotional state of the user has

to be produced. Taking into account current affective cues, responses have to be prepared already before the user has finished speaking and hypotheses have to be generated on-the-fly. It might not make sense at first to detect emotion from speech every few milliseconds, since emotion remains bound to syllables or even words [31]. However, if the output is smoothed over short time periods, reliable estimates of quasi-instantaneous emotions can be given, without the need to identify word boundaries, for example. Moreover, if word boundaries are known, e.g. from an Automatic Speech Recognition (ASR) unit, the emotion output can be mapped to the word segments and averaged to transcribe the emotional content of every word instantaneously.

In the context of emotion-related virtual agents Long Short-Term Memory models have been suggested [22]. Stemming from a similar motivation, our on-line speech-based emotion recognition approach is based on Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN). Those networks directly operate on short time frames of low-level audio descriptors such as Energy, Pitch, and Mel-Frequency-Cepstral Coefficients (MFCC) and do not require any segmentation besides the low-level framing. Additionally, linguistic features are incorporated via early fusion. Due to the regression output in an emotion space spanned by valence and activation (as in [4, 15, 44], for example), the approach is no more limited to detection of discrete emotion classes. For every low-level audio input frame an output value ranging from -1 to $+1$ for each emotional dimension is generated.

This article is structured as follows: the next section addresses the problem of continuous emotion recognition in more detail. In Sect. 3 we briefly outline the SEMAINE architecture, which reveals the structure of a virtual agent system demanding for incremental emotion recognition. Section 4 describes the acoustic and linguistic features which were extracted from the audio signal. Section 5 presents and discusses the propagated LSTM-RNN architecture as well as other Neural Network architectures, to which we compare the performance of the LSTM-RNN. Section 6 introduces the naturalistic emotion database used for experimental evaluations in Sect. 7. Section 8 shows results obtained on this database. We conclude the article in Sect. 9 with an outlook for future trends in fully continuous emotion recognition.

2 The challenge of continuous emotion recognition

As opposed to speech recognition, emotion recognition from single 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. Typical units of analysis are complete sentences, sentence fragments (such as syntactical chunks), or words [38]. The term 'fragment' 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 [28, 31, 32].

Most approaches to emotion recognition model the long range dependencies between low-level signal frames on the feature level. Various characteristics (e.g. statistical functionals) are computed from the temporal contours of low-level audio features. The variable length contours are mapped to a single high-dimensional feature vector for each input fragment. Both classification (for emotion classes) and regression (emotion dimensions) tasks can be solved using this approach, given suitable models, e.g. Support-Vector-Machines for classification and Support-Vector-Regression for regression. A major drawback of these approaches is that prior segmentation is required and usually a complete input fragment is required for analysis. Further, only one output can be produced at the end of an input fragment.

Thus, for continuous output at a fixed rate in the sub-second region, the segment length has to be very small. In this case the amount of context considered is very limited, which impacts recognition performance. Supra-segmental modelling techniques, for example, perform well on the utterance level, due to effects like a highly informative pitch and energy contour throughout the utterance, which is correlated to the conveyed emotion. Randomly chosen sub-second segments, however, do not contain such information, because—without additional effort—we do not know what part of the utterance they are from.

An alternative approach is to use Hidden Markov Models with frame-wise low-level descriptors as observation features [29, 41]. However these approaches also require some kind of segmentation, because they assign a class to each segment. Principally it would be possible to do some kind of time-continuous decoding with such an approach, i.e. as is done for continuous automatic speech recognition. However, this technique is limited to discrete classes of emotion or affect. Thus, these approaches are inherently unsuited for emotion recognition in continuous dimensions.

For fully continuous emotion recognition we must abandon the requirement of defining a suitable unit of analysis, within which the emotional state is assumed as quasi-stationary. As features, only frame-based features must be used, the long range dependencies must be modelled by the classifier, and ideally an output of the current state should be generated for every input frame. In Sect. 5 we will present a classifier which meets all these requirements.

The next section will briefly explain the architecture of an affect aware virtual agent created by the SEMAINE project, where the proposed emotion recognition approach will be used.

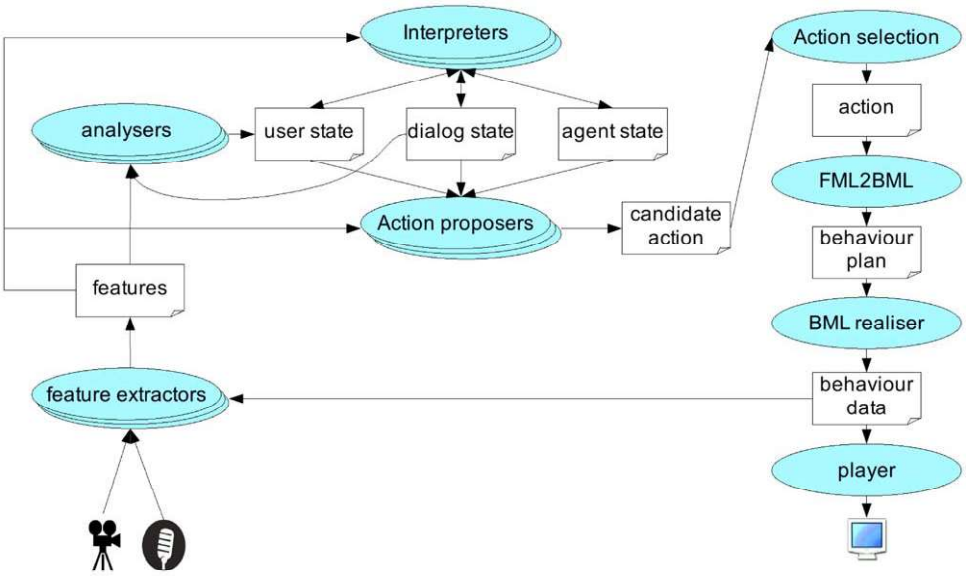


Fig. 1 Architecture of the SEMAINE system

3 The SEMAINE system architecture

The aim of the SEMAINE project[1] is to build a Sensitive Artificial Listener—a multimodal dialogue system with the social interaction skills needed for a sustained conversation with a human user. Figure 1 shows the system architecture for a virtual agent as used in the SEMAINE system [27]. As mentioned before, this system demands for on-line incremental emotion recognition in order to select responses as early as possible. In Fig. 1 processing components such as the LSTM-RNN, which will be outlined in Sect. 5, or the feature extractor (see Sect. 4) are represented as ovals, whereas rectangles denote data. Arrows are always between components and data, and indicate which data is produced by or is accessible to which component. It can be seen that the rough organisation follows the simple tripartition of input (left), central processing (middle), and output (right), and that arrows indicate a rough pipeline for the data flow, from input analysis via central processing to output generation. In particular for the emotion coding, EmotionML is used [26] which already allows for continuous spatio-temporal emotion representation.

The architecture is deliberately held very open and flexible, since the system is intended as a research platform. Thus it allows researchers to easily integrate and test their own components in the system.

The main aspects of the architecture in general are outlined in the following: *feature extractors* analyse low-level audio and video signals, and provide feature vectors (see Sect. 4 for details on audio features) periodically (10 ms) to the analysers which process the low-level features and produce e.g. a representation of the current user state, in terms of e.g. epistemic-affective states (emotion, interest, etc.), or a representation of the user's facial state in terms of facial action units. Another typical output of an analyser is the spoken word chain, or keywords uttered by the user. Since, e.g. automatic speech recognition or emotion recognition might benefit from the dialogue context or user profiles at a higher level, *interpreter* components are contained in the system to address this issue. Such components evaluate the user state in the context of the current state of information regarding the user (including input such as emotion or words obtained from the analysers), the dialog, and the agent itself, and update these information states. Interpreter components include the turn-taking interpreter, for example, which uses low-level audio/video features such as energy, pitch, presence of face, and direction of view, for example, along with user state information to determine whether the user has a turn or the agent should take the turn. Another interpreter component is the utterance interpreter. It analyses the words uttered by the user and assigns function related attributes like agree/disagree, positive/negative, etc. to the words and utterances.

In the future an interpreter for choosing the most likely ASR hypothesis based on the current dialogue state could be added, thereby creating a link between the dialogue state and the user state.

The next group of components is a set of *action proposers* which produce agent action candidates independently from one another. The action proposers take their input mainly from the user, dialog, and agent state. To allow for a flexible architecture a link from the features to the action proposers is included in the architecture for future use. The current set of action proposers included the following components: an utterance producer which will propose the agent's next verbal utterance, given the dialog history, the user's emotion,

[1] http://www.semaine-project.eu/

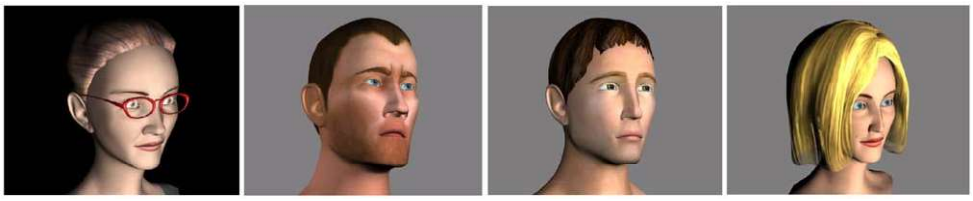


Fig. 2 Faces of virtual characters used in the SEMAINE system (*left* to *right*): Prudence, Spike, Obadiah, and Poppy

the topic under discussion, and the agent's own emotion. An automatic backchannel generator identifies suitable points in time to emit a backchannel (head nod, or vocalisation, for example). A mimicry component will propose to imitate, to some extent, the user's low-level behaviour. Finally, a non-verbal behaviour component needs to continuously generate some 'background' behaviour, especially when the agent is listening but also when it is speaking. The actions proposed may be contradictory, and thus must be filtered by an *action selection* component. A selected action is converted from a description in terms of its functions into a behaviour plan, which is then realised in terms of low-level data that can be used directly by a player.

The SEMAINE system allows the user to choose between four different virtual agent characters to talk to. These are shown in Fig. 2 and are detailed in [7]. Each character has a different 'personality': 'Prudence' is matter-of-fact, 'Spike' is aggressive, 'Obadiah' is pessimistic, and 'Poppy' is cheerful. The character selection in the current system can be performed manually by clicking on the character's face in the control GUI, or far more naturally by simply telling the character one is currently talking to, that one wishes to talk to someone else. In certain cases where the system detects problems in the conversation, e.g. due to a bored user or highly aroused/annoyed user, it might by itself ask the user if he or she wants to talk to somebody else.

4 Features

Despite the finding that multimodal systems which also include vision-based features might lead to better recognition performance [4, 5, 33, 45], this article exclusively focuses on audio features. Yet, in principle our approach allows for the inclusion features from the video channel or other modalities, provided that early fusion is applied. Acoustic features from the speech signal were extracted using the openSMILE feature extractor [8], which was also used to provide features for the Interspeech 2009 Emotion Challenge [34]. Further, a keyword spotter generates linguistic features that are processed frame-wisely using early fusion.

The following two sub-sections describe the acoustic and linguistic features which are used for emotion recognition in detail.

Table 1 28 low-level audio features for the propagated frame-based emotion analysis (column C) and 39 features for turn-based (supra-segmental) reference evaluations (column T)

Feature Group	Features in Group	C	T
Signal energy	Root mean-square and log. energy	1	2
Pitch	Fundamental frequency F_0, 2 measures for probability of voicing	1	3
Voice quality	Harmonics-to-noise ratio	1	1
Cepstral	MFCC	12	16
Time signal	Zero-crossing-rate, max. and min. value, DC component	1	4
Spectral	Energy in bands 0–250 Hz, 0–650 Hz, 250–650 Hz, 1000–4000 Hz	4	4
	10%, 25%, 50%, 75%, and 90% roll-off	5	5
	Centroid, flux, and relative position of maximum and minimum	3	4
SUM:		28	39

4.1 Acoustic features

As opposed to previous work in the field of emotion recognition (e.g. [32, 44]), the propagated approach in this work does not rely on supra-segmental modelling based on statistical functionals applied to low-level audio descriptors. The featured Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN) are capable of modelling all necessary long-range dependencies, as will be explained in Sect. 5. The 28 low-level descriptors (LLD) extracted from the audio signal for fully continuous emotion recognition are summarised in Table 1 (column 'C'). The descriptors were extracted every 20 ms for overlapping frames with a frame-length of 32 ms. First order delta coefficients are appended to the 28 LLD, resulting in a $28 \cdot 2 = 56$ dimensional feature vector for each frame.

On an AMD Phenom 64 bit CPU at 2.2 GHz, the openSMILE feature extraction module [8], which is used in our frame-based recognition system, runs on-line with a real-time factor (RTF) of 0.01.

In order to compare results of time-continuous emotion recognition to the turn-based emotion recognition task from [44], a traditional, large, and open set of features is

Table 2 36 statistical functionals applied to the low-level descriptor contours for turn-based emotion analysis

Functionals	#
Maximum/minimum value and relative position	4
Range (max.-min.)	1
Mean and mean of absolute values	2
Max.-mean, min.-mean	2
Quartiles and inter-quartile ranges	6
95%, and 98% percentile	2
Std. deviation, variance, kurtosis, skewness	4
Centroid of contour	1
Linear regression coefficients and approximation error	4
Quadratic regression coefficients and approximation error	5
Zero-crossing rate	1
25% down-level time, 75% up-level time, rise-time, fall-time	4

generated by applying statistical functionals to low-level-descriptor contours. An extended set of 39 low-level descriptors (detailed in Table 1, column 'T') is extracted, first and second order delta coefficients are appended, and 36 functionals are applied to each of the resulting 117 low-level descriptor contours. This results in a total of 4 212 turn-based features. The 36 functionals are described in Table 2.

The 4 212 features for turn-based emotion recognition are reduced to relevant features for activation and valence independently by a Correlation based Feature Subset selection (CFS) [43]. 60 features for activation and 64 features for valence are thereby automatically selected. Please note that in contrast, continuous (frame-based) emotion recognition with LSTM-RNN uses the full set of $28 \cdot 2 = 56$ features (28 acoustic low-level descriptors with 28 delta coefficients appended) without further reduction by feature selection.

All features (turn-based functionals and low-level features) are standardised to have zero mean and unit standard deviation. Both, means and variances are computed from the training data only and are applied for normalising training and test data. Therefore the standardisation can also be performed in an on-line recognition application.

4.2 Linguistic features

Knowledge about the spoken content is incorporated at the frame level via early fusion. The 56 dimensional low-level acoustic feature vector (28 LLD with delta coefficients appended, see Sect. 4.1) is extended by appending N_l binary linguistic features. Thereby each binary feature corresponds to the occurrence of one of 56 keywords that were shown to be correlated to either valence, activation, or both. The keywords were selected out of a lexicon of size 1 915 by applying CFS on the training set. Thereby 21 out of 56 keywords were found to be correlated to activation, while 40 were found to be correlated to valence. Thus, $N_l = 21$ binary linguistic features are append to the acoustic features for activation classification tasks and $N_l = 40$ for valence classification. Keywords like *again*, *angry*, *assertive*, *very* etc. were selected for activation, and typical keywords correlated to valence where e.g. *good*, *great*, *lovely*, or *totally*. The 56 keywords consisted of nouns (14%), adverbs/adjectives (30%), verbs (18%), and others such as pronouns or prepositions (38%).

Note that using a single linguistic feature containing the current word identity in form of a word index would not be feasible with LSTM networks since they assume that the absolute value of a feature is always correlated or proportional to the 'intensity' of the corresponding feature. This, however, would not be true for a 'word index feature'. Our technique is related to Bag of Words modelling [18]. However, since our approach bases on incremental processing, only one keyword can occur at a given time frame.

For combined acoustic-linguistic analysis, a short buffer (1–4 seconds, depending on keyword length and desired accuracy) has to be included in order to allow the keyword spotter to provide the binary features *after* the keyword has been decoded. Yet, this causes only a short delay and does not contradict our principle of on-line recognition because linguistic features can be delivered (although slightly delayed) while the user is speaking.

The keyword spotter was trained on the TIMIT database and re-trained on the training split of the Sensitive Artificial Listener database (see Sect. 6) to allow a better modelling of emotionally coloured and spontaneous speech (see [48] for a detailed description of the keyword spotter model architecture and parameterisation).

A true positive rate of 58.5% at a false positive rate of 5.5% is achieved on the SAL test set by the recogniser used within the presented system. This shows the difficulty of detecting (partly very short) keywords in emotionally coloured spontaneous speech (for comparison: [48] reports 97.5% true positive rate on TIMIT at a false positive rate of under 1.4% using the same approach).

5 LSTM-RNN

As a well suited technique for on-line regression of emotion dimensions we applied a specialised neural network architecture called Long Short-Term Memory RNN [17]. Traditional feed-forward neural networks such as the multi-layer perceptron are not suitable for classification of connected time series, as they are static classifiers which classify data frame by frame without considering neighbouring 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

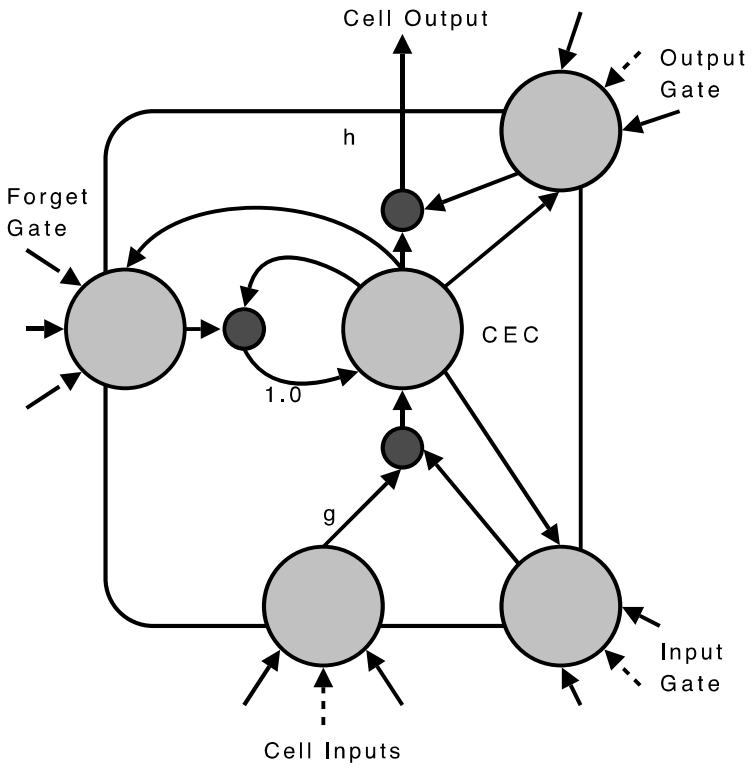


Fig. 3 LSTM memory block consisting of one memory cell: the input, output, and forget gates collect activations from inside and outside the block which control the cell through multiplicative units (depicted as *small circles*); input, output, and forget gate scale input, output, and internal state respectively;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

itself. Thus, the network can learn to model past events by adjusting the weights of the feed-back connection(s).

However, 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). This lead to various attempts to address the problem of vanishing gradients for RNN, including non-gradient based training [2], time-delay networks [19, 20, 24], and hierarchical sequence compression [25]. One of the most effective techniques is the Long Short-Term Memory architecture [17], which is able to store information in linear memory cells over a longer period of time. LSTM architectures have shown good performance in many tasks for which context modelling is essential, e.g. phoneme recognition [12], keyword spotting [9, 46], handwriting recognition [14, 21], noise modelling [49], and emotion recognition from speech [44, 47]. An LSTM layer is composed of recurrently connected memory blocks, each of which contains one or more recurrently connected memory cells (cf. Fig. 3), 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 recognise emotion on a frame level. A more detailed explanation of the LSTM principle can be found in [11].

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 LSTM (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 Fig. 4), resulting in Bidirectional Long Short-Term Memory Recurrent Neural Networks (BLSTM-RNN) [36]. This method is applied especially for speech recognition tasks [10, 12], to model anticipatory coarticulation effects. Thereby each hidden layer is duplicated, while one layer processes the inputs forwards and the other backwards.

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 on-line classification. Therefore, we will not put our primary focus on this architecture,

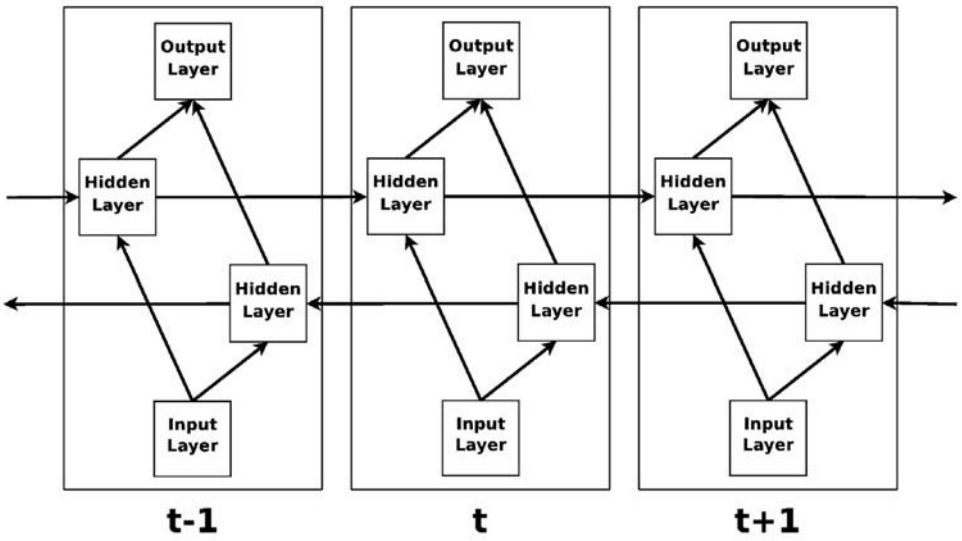


Fig. 4 Bidirectional Recurrent Neural Network

even though we will show by a few exemplary results, that this bidirectional network architecture yields better results than the unidirectional architecture (Sect. 8).

LSTM-RNN and BLSTM-RNN can both be trained via standard backpropagation through time (BPTT) [42]. A variant of the standard backpropagation algorithm is Resilient Propagation (rProp) [23], where only the sign of the error gradient is considered for network weight updates, not the absolute value multiplied by a learn rate parameter. 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. Thus, resilient propagation is used in our evaluations. More details on the configurations of the specific networks used for evaluations within this work can be found in Sect. 7.

In contrast to the BLSTM-RNN which requires future speech frames, and is therefore more suited for off-line processing, the 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.

6 Database

Many small to mid-size emotional speech databases exist, where emotion is labelled as discrete classes and often only emotional prototypes are contained, e.g. [3] or [39]. Further, there are databases in which emotion is labelled turn-wise on a continuous valence-activation scale such as the Vera am Mittag (VAM) corpus [16]. For continuous recognition of emotion in time as well as in a 2-D space spanned by activation and valence, new databases and different annotation techniques are required. Such an annotation tool is FEELtrace, which was introduced in [6]. The first, and so far only, naturalistic database annotated using the FEELtrace system, which is made available to researchers is the Belfast naturalistic sub-set of the HUMAINE database [7]. This sub-set is also known as the Belfast Sensitive Artificial Listener (SAL) data and will be used for all experiments in this article.

The database contains 25 audio-visual recordings in total from four speakers (two male, two female) with an average recording length of 20 minutes per recording session and 2 recording sessions per speaker. The recordings were obtained during natural human-computer conversations, which were recorded using a Wizard-of-Oz SAL interface designed to let users work through a range of emotional states. The database and the recording procedure is described in more detail in [7]. Data was labelled with respect to the emotional dimensions valence and activation by four annotators continuously in real-time while listening to the recordings using the FEELtrace system. The adjusted values for valence and activation were sampled every 10 ms to obtain a temporal quasi-continuum. Note that a certain delay is incorporated into the annotations since the labellers require a finite time to react to a change of emotion.

As ground truth the mean of all four annotators was computed. The average Mean Squared Error (MSE) of the four human annotators with respect to the mean value is 0.08 for activation and 0.07 for valence, corresponding to a standard deviation of 0.17 and 0.14, respectively.

For the speaker dependent experiments reported on in this article the same training- and test-set splits as introduced in [44] are used, in order to be able to compare results. Thereby, the 25 sequences are split into 16 training sequences and 9 test sequences. The test split has a total length of 53.3 minutes whereas the training set has a length of 99.2 minutes.

Since only four speakers are contained in this data-set, the training- and test-splits, which we focus on, are not

speaker disjunctive. Speaker dependent emotion recognition is of significant practical importance, especially for the paradigm of virtual agents and sensitive listeners, since the listener can adapt it's models to the current speaker and learn speaker profiles. However, in order to provide informative results for speaker independent performance, we additionally use a speaker independent set (data from two speakers is used for training and data from the other two speakers is used for testing).

To obtain a baseline result by evaluating the LSTM-RNN and the audio features on the same turn-based emotion recognition task as in [44], the 25 recording sequences have been split into turns using an energy based Voice Activity Detection, where a 'turn' corresponds to the audio from the end of one silence segment to the beginning of the next silence segment. A total of 1 692 turns with an average length of 3.5 seconds is accordingly contained in the database. The average number of words per turn is 10.4. Training- and test splits contain 1 102 and 590 turns, respectively. Labels for each turn are computed by averaging the FEELtrace labels for valence and activation over a complete turn.

Apart from the necessity to deal with continuous values for time and emotion, the great challenge of emotion recognition on the naturalistic SAL database is the fact that the system must deal with all data—as observed and recorded—and not only manually pre-selected 'emotional prototypes' as in many other databases. Note that there is usually a high difference in accuracy between the tasks of prototypical and non-prototypical emotion recognition [34, 35, 37]. E.g. on the FAU AIBO corpus [1], which contains real-life non-prototypical emotions, an unweighted average recall rate of approximately 38% is achieved in [34] for a five class problem. For a database of acted prototypical emotions with seven classes [3], an unweighted average recall of approximately 85% is achieved in [35].

7 Evaluation

Four substantially different evaluations are performed:

- Turn-based emotion recognition with acoustic features as a reference evaluation (turn.(A))
- Frame-based (time-continuous) emotion recognition with acoustic features (cont.(A))
- Frame-based (time-continuous) emotion recognition with linguistic features and acoustic and linguistic features combined (cont.(L), cont.(A+L)).

Table 3 summarises the settings for the four evaluations. Note that the BLSTM-RNN have two hidden layers, (one for the forward direction and one for the backward direction), each consisting of 50 memory blocks with one LSTM cell per memory block.

Table 3 Evaluations performed and according configuration. Classifiers: standard Recurrent Neural Networks (RNN), (bidirectional) Long Short-Term Memory RNN ((B)LSTM), and Support Vector Regression (SVR). N_h is the size (number of LSTM memory blocks with one LSTM memory cell each) of the single hidden layer of the LSTM-RNN or each of the two hidden layers of the BLSTM-RNN

Eval.	Classifier	Features	N_h
Turn.(A)	(B)LSTM, RNN, SVR	Acoustic	50
Cont.(A)	(B)LSTM, RNN	Acoustic	50
Cont.(L)	(B)LSTM, RNN	Linguistic	50
Cont.(A+L)	(B)LSTM, RNN	Ac. & ling.	70

As a common continuous recognition technique, regression by Support-Vector-Regression (SVR) is performed for comparison [15, 43, 44]. The Support-Vector-Regression used a polynomial kernel function of degree 1, complexity $C = 1.0$, and Sequential Minimal Optimisation (SMO). In order to show the true benefit of LSTM context modelling, a standard recurrent network with one hidden layer having 50 neurons was furthermore evaluated.

Due to the increased size of the combined acoustic-linguistic feature vector the LSTM network size is increased from 50 memory blocks to 70 memory blocks (cont.(A+L)). For all LSTM and BLSTM evaluations, experiments where static noise with a standard deviation of $\sigma_n = 0.3$ is added to the features of the training data are conducted in addition to those experiments where no static noise was added to the features. This technique is claimed to improve generalisation of the networks and thus can increase performance on unknown data [13]. During training and evaluation all turns were presented to the network in the correct temporal order so that—in case of turn-based recognition—the classifier can make use of (bidirectional) context between *turns*.

The training algorithm for the LSTM and BLSTM networks converged after 100–200 epochs.

The obtained (B)LSTM-RNN predictions $x[n]$ at time n (frame/turn index) are smoothed via first order low-pass filtering using the following equation ($x_s[n]$ denotes the filtered predictions):

$$x_s[n] = \alpha x_s[n-1] + (1-\alpha) \cdot x[n] \quad (1)$$

An α of 0.99 was used for frame-based emotion recognition and an α of 0.7 was used for turn-based recognition.

8 Results and discussion

This section presents and discusses the results obtained for the evaluations described in Sect. 7. All results are given as cross correlation coefficient (CC), measuring the similarity between the annotations and the classifier outputs. Thereby

Table 4 Correlation coefficients (CC) for activation (upper half of the table) and valence (lower half) obtained with Long Short-Term Memory (LSTM), bidirectional LSTM (BLSTM), Support Vector Regression (SVR), and standard Recurrent Neural Networks (RNN); speaker dependent tests; turn-based recognition (turn.) and time-continuous, frame-wise recognition (cont.); static Gaussian noise with standard deviation $\sigma_n = 0.3$ added to training data, and unmodified training data ($\sigma_n = 0$); A: acoustic features, L: linguistic features, A+L: acoustic and linguistic features

Model	σ_n	Turn. (A)	Cont.		
			(A)	(L)	(A+L)
CC for activation (speaker dependent)					
BLSTM	0	0.51	0.46	0.18	0.37
BLSTM	0.3	**0.57**	0.46	0.15	**0.51**
LSTM	0	0.56	**0.47**	0.24	0.46
LSTM	0.3	0.53	0.42	**0.32**	0.13
RNN	0.0	0.40	0.03	0.18	0.02
SVR	0.0	0.30	–	–	–
CC for valence (speaker dependent)					
BLSTM	0	0.37	**0.54**	**0.42**	**0.55**
BLSTM	0.3	0.31	0.51	0.16	0.34
LSTM	0	0.26	0.48	0.42	0.42
LSTM	0.3	**0.37**	0.37	0.32	0.53
RNN	0.0	0.28	0.37	0.14	0.38
SVR	0.0	0.28	–	–	–

a correlation coefficient of 1.0 corresponds to perfect prediction whereas a CC of 0.0 indicates chance level. Note that the linear or quadratic error measures are not the best measure for emotion recognition performance. Rather, the correlation coefficient should be preferred since it measures the ability of the network to follow the labels in general. Offsets and 'noise' in the result do not influence the CC as much as they influence the mean squared error, for example.

The upper part of Table 4 shows the correlation coefficients for the prediction of activation. Best results can be obtained when using a turn-based approach, however, a time-continuous predictor using acoustic features achieves comparable results (a CC of 0.46 and 0.47 for the BLSTM and the LSTM, respectively). In most cases, the BLSTM slightly outperforms the unidirectional LSTM. Turn-based acoustic analysis was also carried out applying SVR which gave significantly poorer results (CC of 0.30). Further evidence that classifiers using long range contextual information prevail over conventional predictors was collected when evaluating standard RNNs: the turn-based prediction of activation using an RNN as described in Sect. 7 with acoustic features leads to a CC of only 0.40. In case of continuous classification the CC of RNN prediction is as low as 0.03 which is close to the chance level of 0. Thus, the RNN is not able to model the long range context necessary for the frame-based activation recognition, while in contrast it is comparably well suited for turn-wise recognition. The short-term context captured by the RNN seems to be beneficial here (RNN outperform SVR), as adjacent turns seem to have similar emotion labels.

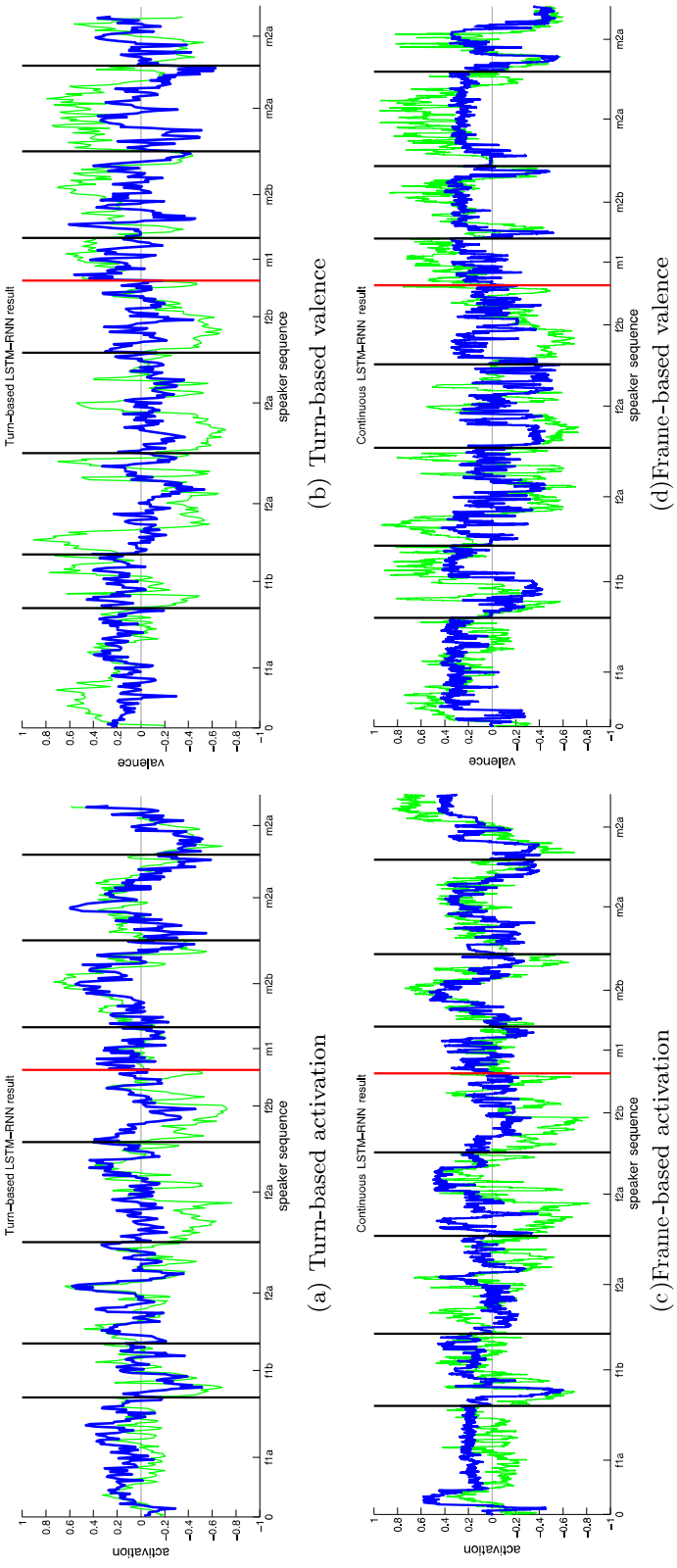
In contrast to the prediction of activation, the recognition of valence profits from frame-based models as can be seen in the lower part of Table 4. A possible explanation might be that valence is conveyed more by spectral and linguistic cues, while activation is conveyed more by prosody. Thus, prosody for activation is captured well by supra-segmental modelling. However for valence, longer terms might smear spectral and linguistic information and thus supra-segmental modelling on a high level is not suited well. Smaller segment sizes, e.g. words as investigated in [28] might improve the result for supra-segmental modelling. This finding is undermined by the fact that standard RNNs perform much better for valence (acoustic features) than for activation (CC 0.37 and 0.03, respectively), which shows that valence information is conveyed more locally, while activation information is conveyed over a longer period of time.

Please note further, that valence is generally harder to determine from the acoustic signal than activation [44]. When exclusively using acoustic features, the continuous approach outperforms the corresponding turn-based RNNs. A further performance gain can be achieved when adding linguistic features, resulting in a CC of 0.53 and 0.55 for the LSTM and the BLSTM, respectively. Again, SVR can not compete with context sensitive LSTM networks (CC of 0.28 using only acoustic features).

Comparing performances for acoustic and linguistic features separately, we see that acoustic features generally are superior to linguistic ones, however, the combination of both slightly improves performance in some cases and decreases performance in other settings. No clear trend is visible, except that the overall best result for each dimension is most often achieved with acoustic and linguistic features combined. A notable exception is the performance of CC = 0.13 for activation with LSTM and combined features (upper part of Table 4). Obviously, in this case the training algorithm converged in a local minimum and was aborted too early. Note that during training we used the root mean square (RMS) error as target function while we report the correlation coefficient as performance measure. Optimising with respect to the correlation coefficient could therefore be an interesting alternative for future investigations in order to improve training stability.

Figures 5(a) and 5(b) visualise the prediction quality of turn-based LSTM networks. Both, the annotation and the LSTM prediction are plotted over time. Obviously, activation (Fig. 5(a)) is predicted more accurately.

The prediction of activation and valence using a continuous LSTM is shown in Figs. 5(c) and 5(d). Still the estimation of valence remains challenging, however, when comparing Figs. 5(d) and 5(b), the benefit of time-continuous



(a) Turn-based activation (b) Turn-based valence

(c)Frame-based activation (d)Frame-based valence

Fig. 5 Prediction for activation/valence using LSTM (*dark*, *thick line*; online: *blue*); ground truth (*light*, *thin line*; online: *green*) for all turns/frames of the test set in sequence (time in turns/frames on the abscissa). *Vertical bars* separate the speaker sequences, i.e. data from one speaker recorded in one continuous recording session (cf. Sect. 6); speaker sequences contain a speaker ID consisting of gender (m/f) and a number followed by a letter indicating the index recording sequence from this speaker (e.g. 'f1a', referring to the first recording sequence of the first female speaker in the test set). No static noise, acoustic features only

Table 5 Correlation coefficients (CC) for activation (act.) and valence (val.) obtained with Long Short-Term Memory (LSTM), and bidirectional LSTM (BLSTM); speaker independent tests; for turn-based recognition (turn.) and time-continuous, frame-wise recognition (cont.); static Gaussian noise with standard deviation $\sigma_n = 0.3$ added to training data, and unmodified training data ($\sigma_n = 0$); acoustic features only

Model	σ_n	Turn.		Cont.	
		Act.	Val.	Act.	Val.
CC for val./act. (speaker independent)					
BLSTM	0	0.44	−0.05	0.26	0.03
BLSTM	0.3	**0.46**	**0.14**	**0.53**	**0.39**
LSTM	0	0.34	0.06	0.14	0.16
LSTM	0.3	0.30	−0.26	0.25	0.32

valence prediction can be retraced for many speech segments. Note that partly a better correlation coefficient could be obtained when scaling the BLSTM output activations by approximately 1.3. Yet, we decided to display the raw outputs in order to give an impression of the intrinsic prediction quality without any post-processing.

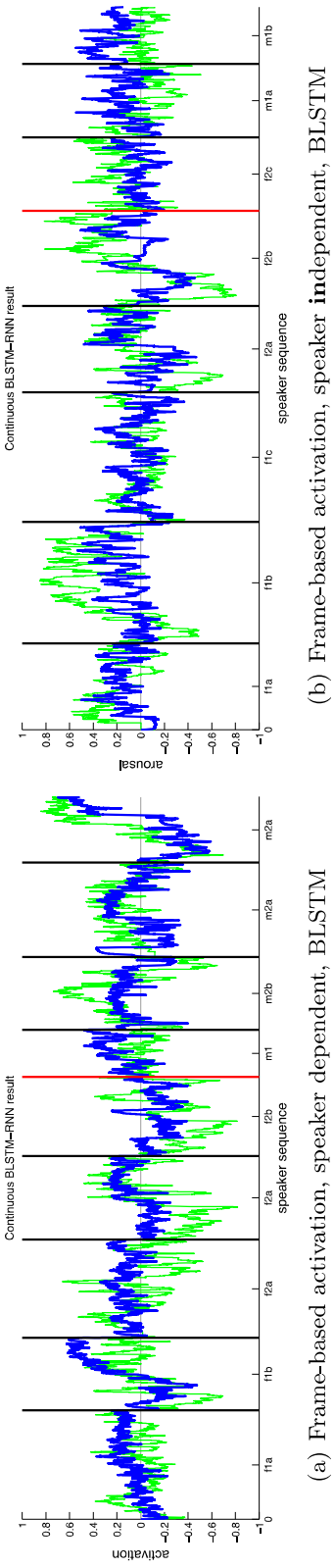
Finally, we provide speaker independent recognition results in Table 5 and a plot comparing speaker independent recognition (Fig. 6(b)) with speaker dependent recognition (Fig. 6(a)) of activation. An interesting observation can be made here: adding static noise to the training data significantly improves results in the speaker independent test for both activation and valence, thus undermining the theory that a better generalisation is achieved by this technique (see Sect. 7).

9 Conclusion and outlook

This article presented a framework for fully time-continuous affect recognition in an emotional space spanned by activation, valence, *and time*. Since our system operates in real-time, it can be applied for virtual agents as the described SEMAINE system, which require an incremental prediction of the user's emotion at every time frame. Our approach can be seen as a first step towards closing the gap between speech recognition—where incremental processing is already state-of-the-art—and emotion recognition, which so far mostly focuses on classifying predefined speech segments only at the end of a spoken utterance.

For the prediction of valence, which is known to be the most challenging emotional dimension, our time-continuous recognition architecture even outperforms a turn-based approach. A further performance gain was achieved by including linguistic features.

Of course the inclusion of other emotional dimensions (such as dominance or control) into our framework is also



(a) Frame-based activation, speaker dependent, BLSTM

(b) Frame-based activation, speaker **in**dependent, BLSTM

Fig. 6 Prediction for activation using BLSTM (*dark, thick line*; online: *blue*), comparison of speaker dependent and speaker independent models; ground truth (*light, thin line*; online: *green*) for all turns/frames of the test set in sequence (time in turns/frames on the abscissa). *Vertical bars* separate the speaker sequences, i.e. data from one speaker recorded in one continuous recording session (cf. Sect. 6); speaker sequences contain a speaker ID consisting of gender (m/f) and a number followed by a letter indicating the index recording sequence from this speaker (e.g. 'f1a', referring to the first recording sequence of the first female speaker in the test set). No static noise, acoustic features only

possible, provided that a sufficient correlation between the dimension of interest and the acoustic/linguistic features exists. An important post-processing step which deserves further investigation is the quantisation of continuous values for the emotional dimensions before passing the classifier output to the dialogue management, e.g. by clustering points in the emotional space [47].

Future works will focus on investigating the benefit of including further feature types such as different kinds of linguistic features and vision features into a time-continuous context sensitive emotion recognition framework. Also the LSTM architecture and parameterisation could be improved by including more hidden layers, using different layer sizes, and especially training on larger databases, once available.

Since we observed improved results for bidirectional LSTM networks, which however are difficult to implement in a causal real-time framework, the investigation of the potential of BLSTM-RNN for on-line recognition is exceedingly promising. A possible approach would be a Tandem system with an LSTM-RNN that produces immediate outputs which are refined over time by a BLSTM as more frames become available.

Even though the amount of social competence our emotion recognition framework can incorporate into a virtual agent remains limited and cannot fully compete with human affect recognition quality and power of observation, three important preconditions for future conversational agents are met: the inclusion of contextual information, the possibility to deliver an estimate of the user's affect at every time frame, and operation in real-time.

Acknowledgements 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), and from the WinReNN project granted by the Bundesministerium für Umwelt.

References

1. Batliner A, Steidl S, Nöth E (2008) Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus. In: Deviller L, Martin JC, Cowie R, Douglas-Cowie E, Batliner A (eds) Proc. of a satellite workshop of LREC 2008 on corpora for research on emotion and affect, pp 28–31. Marrakesh
2. Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw 5(2):157–166
3. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A database of German emotional speech. In: Proc. of interspeech, pp 1517–1520. Lisbon, Portugal
4. Caridakis G, Malatesta L, Kessous L, Amir N, Raouzaiou A, Karpouzis K (2006) Modeling naturalistic affective states via facial and vocal expressions recognition. In: Proc. of the 8th international conference on multimodal interfaces, pp 146–154. Banff, Alberta, Canada,
5. Castellano G, Kessous L, Caridakis G (2008) Emotion recognition through multiple modalities: face, body gesture, speech. In: Peter C, Beale R (eds) Affect and emotion in human-computer interaction. Springer, Berlin, pp 92–103
6. Cowie R, Douglas-Cowie E, Savvidou S, McMahon E, Sawey M, Schröder M (2000) Feeltrace: an instrument for recording perceived emotion in real time. In: Proceedings of the ISCA workshop on speech and emotion, pp 19–24
7. Douglas-Cowie E, Cowie R, Sneddon I, Cox C, Lowry O, McRorie M, Martin JC, Devillers L, Abrilian S, Batliner A, Amir N, Karpouzis K (2007) The HUMAINE database. In: Proc. of ACII, pp 488–500
8. Eyben F, Wöllmer M, Schuller B (2009) openEAR—introducing the Munich Open-source Emotion and Affect Recognition Toolkit. In: Proc. of ACII, pp 576–581. Amsterdam, The Netherlands
9. Fernandez S, Graves A, Schmidhuber J (2007) An application of recurrent neural networks to discriminative keyword spotting. In: Proc. of ICANN, pp 220–229. Porto, Portugal
10. Fernandez S, Graves A, Schmidhuber J (2008) Phoneme recognition in TIMIT with BLSTM-CTC. Tech. rep., IDSIA
11. Graves A (2008) Supervised sequence labelling with recurrent neural networks. Ph.D. thesis, Technische Universität München
12. Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw 18(5–6):602–610
13. Graves A, Fernandez S, Schmidhuber J (2005) Bidirectional LSTM networks for improved phoneme classification and recognition. In: Proceedings of ICANN, vol 18. Warsaw, Poland, pp 602–610
14. Graves A, Fernandez S, Liwicki M, Bunke H, Schmidhuber J (2008) Unconstrained online handwriting recognition with recurrent neural networks. Adv Neural Inf Process Syst
15. Grimm M, Kroschel K, Narayanan S (2007) Support vector regression for automatic recognition of spontaneous emotions in speech. In: Proc. of ICASSP, pp 1085–1088
16. Grimm M, Kroschel K, Narayanan S (2008) The vera am mittag german audio-visual emotional speech database. In: Proc. of ICME, pp 865–868. Hannover, Germany
17. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
18. Joachims T (1998) Text categorization with support vector machines: learning with many relevant features. In: Proc. of ECML, pp 137–142. Chemniz, Germany
19. Lang KJ, Waibel AH, Hinton GE (1990) A time-delay neural network architecture for isolated word recognition. Neural Netw 3(1):23–43
20. Lin T, Horne BG, Tino P, Giles CL (1996) Learning long-term dependencies in NARX recurrent neural networks. IEEE Trans Neural Netw 7(6):1329–1338
21. Liwicki M, Graves A, Fernandez S, Bunke H, Schmidhuber J (2007) A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In: Proc. of ICDAR, pp 367–371. Curitiba, Brazil
22. Peters C, O'Sullivan C (2002) Synthetic vision and memory for autonomous virtual humans. Comput Graph Forum 21(4):743–753
23. Riedmiller M, Braun H (1993) A direct adaptive method for faster backpropagation learning: the RPROP algorithm. In: IEEE international conference on neural networks, pp 586–591
24. Schaefer AM, Udluft S, Zimmermann HG (2008) Learning long-term dependencies with recurrent neural networks. Neurocomputing 71(13–15):2481–2488
25. Schmidhuber J (1992) Learning complex extended sequences using the principle of history compression. Neural Comput 4(2):234–242
26. Schröder M, Devillers L, Karpouzis K, Martin JC, Pelachaud C, Peter C, Pirker H, Schuller B, Tao J, Wilson I (2007) What should

a generic emotion markup language be able to represent? In: Paiva A, Prada R, Picard RW (eds) Affective computing and intelligent interaction. Springer, Berlin, pp 440–451
27. Schröder M, Cowie R, Heylen D, Pantic M, Pelachaud C, Schuller B (2008) Towards responsive sensitive artificial listeners. In: Proc. of 4th intern. workshop on human-computer conversation. Bellagio, Italy
28. Schuller B, Rigoll G (2006) Timing levels in segment-based speech emotion recognition. In: Proc. of interspeech, pp 1818–1821. Pittsburgh, PA, USA
29. Schuller B, Rigoll G, Lang M (2003) Hidden Markov model-based speech emotion recognition. In: Proc. of ICASSP, pp 1–4. Hong Kong, China
30. Schuller B, Reiter S, Rigoll G (2006) Evolutionary feature generation in speech emotion recognition. In: Proc. of ICME, pp 5–8. Toronto, Canada
31. Schuller B, Vlasenko B, Minguez R, Rigoll G, Wendemuth A (2007) Comparing one and two-stage acoustic modeling in the recognition of emotion in speech. In: Proc. of ASRU, pp 596–600. Kyoto, Japan
32. Schuller B, Wimmer M, Mösenlechner L, Kern C, Arsic D, Rigoll G (2008) Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space? In: Proc. of ICASSP, pp 4501–4504. Las Vegas, Nevada, USA
33. Schuller B, Müller R, Eyben F, Gast J, Hörnler B, Wöllmer M, Rigoll G, Höthker A, Konosu H (2009) Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. Image Vis Comput J 27(12):1760–1774. Special issue on visual and multimodal analysis of human spontaneous behavior
34. Schuller B, Steidl S, Batliner A (2009) The Interspeech 2009 emotion challenge. In: Proc. of interspeech, pp 312–315. Brighton, UK
35. Schuller B, Vlasenko B, Eyben F, Rigoll G, Wendemuth A (2009) Acoustic emotion recognition: A benchmark comparison of performances. In: Proc. of ASRU 2009. Merano, Italy
36. Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Proc 45:2673–2681
37. Seppi D, Batliner A, Schuller B, Steidl S, Vogt T, Wagner J, Devillers L, Vidrascu L, Amir N, Aharonson V (2008) Patterns, prototypes, performance: classifying emotional user states. In: Proc. of interspeech, pp 601–604. Brisbane, Australia
38. Steidl S (2009) Automatic classification of emotion-related user states in spontaneous children's speech. Logos, Berlin
39. Steininger S, Schiel F, Dioubina O, Raubold S (2002) Development of user-state conventions for the multimodal corpus in smartkom. In: Workshop on multimodal resources and multimodal systems evaluation, pp 33–37. Las Palmas
40. Streit M, Batliner A, Portele T (2006) Emotions analysis and emotion-handling subdialogues. In: Wahlster W (ed) SmartKom: foundations of multimodal dialogue systems. Springer, Berlin, pp 317–332
41. Vlasenko B, Schuller B, Wendemuth A, Rigoll G (2007) Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing. In: Paiva A (ed) Proc. of ACII, pp 139–147. Lisbon, Portugal
42. Werbos P (1990) Backpropagation through time: What it does and how to do it. Proc IEEE 78:1550–1560
43. Witten IH, Frank E (2005) Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco,
44. Wöllmer M, Eyben F, Reiter S, Schuller B, Cox C, Douglas-Cowie E, Cowie R (2008) Abandoning emotion classes—towards continuous emotion recognition with modelling of long-range dependencies. In: Proc. of interspeech, pp 597–600. Brisbane, Australia
45. Wöllmer M, Al-Hames M, Eyben F, Schuller B, Rigoll G (2009) A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams. Neurocomputing 73:366–380
46. Wöllmer M, Eyben F, Keshet J, Graves A, Schuller B, Rigoll G (2009) Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks. In: Proc. of ICASSP, pp 3949–3952. Taipei, Taiwan
47. Wöllmer M, Eyben F, Schuller B, Douglas-Cowie E, Cowie R (2009) Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks. In: Proc. of interspeech, pp 1595–1598. Brighton, UK
48. Wöllmer M, Eyben F, Schuller B, Rigoll G (2009) Robust vocabulary independent keyword spotting with graphical models. In: Proc. of ASRU 2009. Merano, Italy
49. Wöllmer M, Eyben F, Schuller B, Sun Y, Moosmayr T, Nguyen-Thien N (2009) Robust in-car spelling recognition—a tandem BLSTM-HMM approach. In: Proc. of interspeech, pp 2507–2510. Brighton, UK
50. Zeng Z, Pantic M, Roisman GI, Huang TS (2009) A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58