Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening

Martin Wöllmer, *Member, IEEE*, Björn Schuller, *Member, IEEE*, Florian Eyben, *Member, IEEE*, and Gerhard Rigoll, *Senior Member, IEEE*

***Abstract*—The automatic estimation of human affect from the speech signal is an important step towards making virtual agents more natural and human-like. In this paper, we present a novel technique for incremental recognition of the user's emotional state as it is applied in a sensitive artificial listener (SAL) system designed for socially competent human–machine communication. Our method is capable of using acoustic, linguistic, as well as long-range contextual information in order to continuously predict the current quadrant in a two-dimensional emotional space spanned by the dimensions valence and activation. The main system components are a hierarchical dynamic Bayesian network (DBN) for detecting linguistic keyword features and long short-term memory (LSTM) recurrent neural networks which model phoneme context and emotional history to predict the affective state of the user. Experimental evaluations on the SAL corpus of non-prototypical real-life emotional speech data consider a number of variants of our recognition framework: continuous emotion estimation from low-level feature frames is evaluated as a new alternative to the common approach of computing statistical functionals of given speech turns. Further performance gains are achieved by discriminatively training LSTM networks and by using bidirectional context information, leading to a quadrant prediction F1-measure of up to 51.3 %, which is only 7.6 % below the average inter-labeler consistency.**

***Index Terms*—Dynamic Bayesian networks (DBNs), emotion recognition, intelligent environments, long short-term memory (LSTM), recurrent neural nets, virtual agents.**

I. Introduction

FOR the design of intelligent environments which enable natural human–machine interaction it is important to consider the principles of interhuman communication as the ideal prototype [1]. While automatic speech recognition (ASR) is already an integral part of most intelligent systems such as virtual agents, in-car interfaces, or mobile phones, a lot more pattern recognition modules are needed to close or at least narrow the gap between the human ability to permanently observe and react to the affective state of the conversational partner in a socially competent way, and the straightforwardness of system responses generated by today's state-of-the-art human–computer interfaces [2], [3]. Therefore, automatic emotion recognition (AER) is an essential precondition to make, e.g., virtual agents more human-like and to increase their acceptance among potential users [4]–[7].

Even though researchers report outstanding recognition accuracies when trying to assign an affective state to an emotionally colored speech turn [8], [9], systems that apply automatic emotion recognition still are only rarely found in every day life. The main reason for this is that emotion recognition performance is often overestimated: apart from examples such as call-center data [10]–[12], databases for interest recognition [13], [14], or other spontaneous speech evaluations [15]–[19], most speech-based AER systems are trained and tested on corpora that contain segmented speech turns with acted, prototypical emotions that are comparatively easy to assign to a set of predefined emotional categories [20]–[22]. Often, only utterances that have been labeled equally by the majority of annotators are used to evaluate AER performance. Yet, these assumptions fail to reflect the conditions a recognition system has to face in real-life usage. Next-generation AER systems must be able to deal with non-prototypical speech data and have to continuously process naturalistic and spontaneous speech as uttered by the user (e.g., as in the Interspeech 2009 Emotion Challenge [23]). More specifically, a real-life emotion recognition engine has to model "everything that comes in," which means it has to use all data as recorded, e.g., for a dialogue system, media retrieval, or surveillance task by using an *open microphone* setting. According to [24], dealing with non-prototypicality is "one of the last barriers prior to integration of emotion recognition from speech into real-life technology."

Thus, in this paper we present and investigate a speech-based system for emotion recognition that is able to cope with spontaneous, non-prototypical, and unsegmented speech. We address the problem of predicting the *quadrant* of an emotional space (spanned by the two dimensions *valence* and *activation*), which best describes the current affective state of the speaker. We willfully omit *dominance* as a further dimension, since we found that activation and dominance are usually strongly correlated. Consequently, the continuum of emotional states is reduced to the four quadrants which can be described as *relaxed/serene* (I), *happy/excited* (II), *sad/bored* (III), and *angry/anxious* (IV) in order to keep the affective state information as simple as

Manuscript received nulldate; revised nulldate; accepted nulldate. Date of publication July 12, 1010; date of current version nulldate. The work was supported by the European Community's Seventh Framework Program (FP7/2007-2013) under Grant 211486 (SEMAINE). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sadaoki Furui.

The authors are with the Institute of Human-Machine-Communication, Technische Universität München, 80333 München, Germany (e-mail: woellmer@tum.de; schuller@tum.de; eyben@tum.de; rigoll@tum.de).

Digital Object Identifier 10.1109/JSTSP.2010.2057200

possible. A further motivation for quadrant quantization of the continuous emotional space is to reduce the multiplicity of possible system responses for the emotion dependent dialogue management of virtual agents, since at some stage, a categorical decision about the user's emotion has to be made before determining a suitable system output. The outlined AER framework is optimized for usage within virtual agent scenarios such as the SEMAINE system for *Sensitive Artificial Listening* [25], which demands for incremental real-time emotion estimation. Applications like the SEMAINE system require customized and immediate feedback based on the emotional state of the user, and responses have to be prepared already before the user has finished speaking. This, however, would hardly be feasible using traditional static classification approaches like support vector machines (SVMs) which classify segmented or fixed length speech segments at the end of a speech turn. Instead, incremental processing demands for techniques that operate on short speech segments while incorporating an adequate and gradually increasing amount of contextual information.

As shown in [26], capturing temporal long-range dependencies is essential for the prediction quality of an AER system and is superior to static SVM modeling. Hence, our technique applies long short-term memory (LSTM) recurrent neural networks [27] which have shown excellent performance in many machine learning applications [28]–[30]. This concept is able to model *emotional history* and overcomes the so-called *vanishing gradient problem* in conventional recurrent neural nets (RNNs). We show that LSTM enables a completely novel approach towards RNN based affect recognition, using low-level features on a frame basis instead of turnwise computed statistical functionals or fixed-length feature vector sequences, as applied in other context-independent RNN systems [31]. Our principle of framewise emotion estimation is related to strategies for speech recognition, where the temporal evolution of low-level descriptors is not only captured by functionals of *features* but by the *classifier*. Such an approach has many advantages: it allows for incremental real-time emotion estimation from speech as it is needed for emotionally sensitive virtual agents and does not need to operate on supra-segmental units of speech (as in almost any other method [32]–[34]). Moreover, the precondition of perfect segmentation is not needed anymore and the AER system can update the emotion prediction *while* the user is speaking. The long short-term memory RNN architecture copes with the fact that speech emotion is a phenomenon observed over a longer time window. Typical units of analysis for static classifiers are complete sentences, sentence fragments (i.e., chunks), or words [35]. Yet, finding the optimal unit of analysis is still an active area of research [9], [36], [37]. Unlike hidden Markov model (HMM)-based methods [38], [39] which also focus on low-level features and perform best-path decoding on the complete input fragment, our technique offers the great advantage that the *amount* of contextual information that is used for emotion recognition is learned during training. In order to refine and update the estimation of a user's emotion once the complete spoken utterance is available, we also investigate the usage of *bidirectional* context [40]. This is done by bidirectional long short-term memory (BLSTM) networks which process the entire speech sequence in forward and backward direction using two hidden layers that are connected to the same output layer. In contrast to the bidirectional system which presumes either offline operation or a short "look-ahead" input buffer, the unidirectional LSTM system can operate in real-time at a moderate computational cost (see Section II.B).

In addition to the acoustic features, the system presented herein also uses linguistic features derived from a dynamic Bayesian network (DBN) for keyword spotting. The DBN is designed in a way that it detects keywords which are correlated to the user's emotion in order to provide a binary linguistic feature vector. In order to also exploit the principle of LSTM modeling for the generation of linguistic features, our system contains an additional LSTM network that provides a discrete phoneme prediction feature to the keyword spotter. This principle of tandem LSTM-DBN modeling was shown to prevail over conventional hidden Markov model-based approaches [41].

The emotion recognition system presented in this paper is trained and evaluated on the Sensitive Artificial Listener (SAL) database [42] which contains natural, spontaneous, and emotionally colored speech. We investigate the accuracy of predicting the quadrants of the emotional space as well as the ability to distinguish high from low activation and valence, respectively. Furthermore, we evaluate the AER performance when considering *neutrality* as a fifth emotional state. We consider both turnwise and framewise classification using BLSTM, LSTM, SVM, and conventional RNN architectures—with and without linguistic features. In addition to continuously estimating valence and activation before assigning the prediction to one of the four quadrants, we also investigate discriminative training on the quadrants.

The rest of this paper is structured as follows. Section II describes the SAL database and gives an overview over the introduced AER system architecture. In Section III, the principle of long short-term memory is introduced. Sections IV and V outline the acoustic and the linguistic feature extractor, respectively. We present experimental results in Section VI and concluding remarks are given in Section VII.

II. Sensitive Artificial Listening

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. This section describes the SAL database which was recorded during a Wizard-of-Oz SAL scenario and will be used in the experimental section of this paper. Further, our AER system architecture will be explained.

A. Database

The SAL corpus is a subset of the HUMAINE database[2] [42] that is continuously labeled in a two-dimensional emotional space spanned by activation and valence. It contains 25 audio-visual recordings in total from four speakers (two male, two female) with an average recording length of 20 minutes per speaker. The language spoken in the database is English.

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

[2]http://emotion-research.net/download/pilot-db/

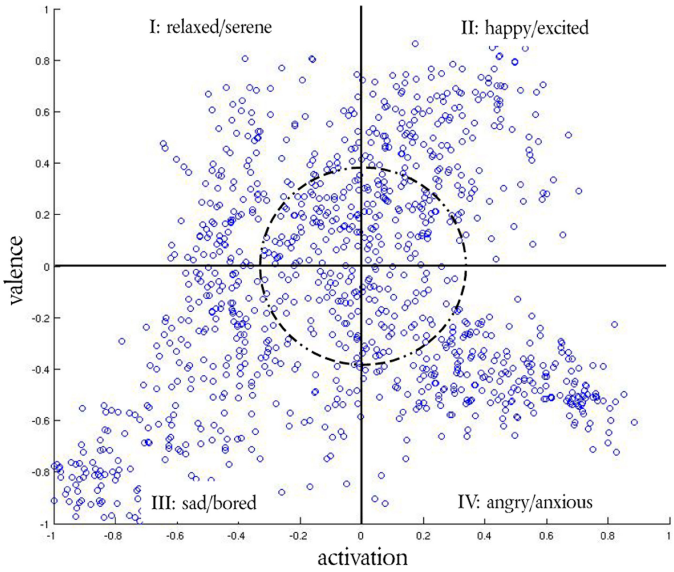

Fig. 1. Turnwise annotations of the SAL database.

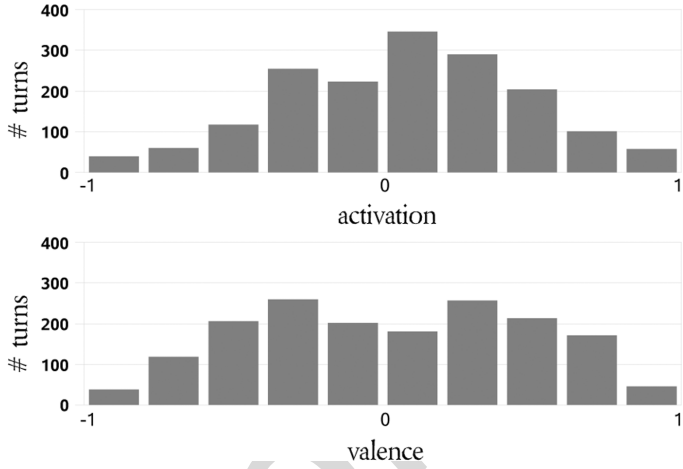

Fig. 2. Histogram for the turnwise annotations of activation (top) and valence (bottom) in the SAL database.

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. All users had to speak to four different virtual characters, each of whom represents one of the four emotional quadrants (Fig. 1): "Prudence" is matter-of-fact (quadrant I), "Poppy" is cheerful (quadrant II), "Obadiah" is pessimistic (quadrant III), and "Spike" is aggressive (quadrant IV). During the conversations, all virtual characters aimed to induce an emotion that corresponds to "their" quadrant. Yet, those "prototypical" virtual characters are used explicitly for emotion induction and not for modeling conditional dependencies between the affective state of the agent and the user, as done in [43] for example. Both, the database and the recording procedure are described in more detail in [42].

The annotators used the FEELtrace system [44] which generates quasi-time-continuous samples of activation and valence every 10 ms (unlike the VAM corpus [45] and practically any other database where labels for the emotional dimensions are given only once per speech turn). All labelers listened to the recordings twice, while annotating activation and valence consecutively in real-time. As ground truth for our experiments, the mean of the four different annotators was used. The mean was calculated by averaging both the (linear) activation and valence coordinates of the labelers for every time step. Note that ambiguous speech turns can lead to the case that the averaged coordinates in the valence-activation space are located in a quadrant that neither of the labelers had assigned to the speech fragment (e.g., the average of coordinates in quadrant I and IV can be located in quadrant II or III). Yet, the resulting quadrant can be seen as the best possible compromise with respect to the average perceived level of activation and valence. An alternative would be to map such ambiguous utterances to a "garbage class." However, since we found that only 2% of the resulting quadrant labels are located in a quadrant that neither of the annotators assigned to the corresponding speech turn, and since *all* of those cases have averaged coordinates that are located in the "neutral" region (coordinates within the dashed circle in Fig. 1), we decided that modeling neutrality is more adequate, rather than the introduction of a "garbage class."

For all experiments reported on in this paper the same training- and test-set splits as introduced in [26] are used. The 25 recording sessions are split into 16 training sessions and nine test sessions. The test split has a total length of 53.3 min, whereas the training split has a length of 99.2 min. Since only four speakers are contained in this database, the training- and test-splits are not speaker disjunctive. Yet, speaker dependent emotion recognition is of significant practical importance, especially for the paradigm of virtual agents and sensitive listeners, since the listener can adapt its models to the current speaker and learn speaker profiles.

For our experiments on turn-based emotion recognition, the sessions were split into turns using an energy based voice activity detection. A total of 1692 turns is accordingly contained in the database. The training- and test splits contain 1102 and 590 turns, respectively. The obtained speech turns do not necessarily comprise complete sentences since the sessions were also split at short hesitation pauses. Thus, the average length of a speech turn is 3.5 seconds. Since the turns are short enough to assume quasi-stationarity of the emotion within a turn, labels for each turn were computed by averaging the FEELtrace annotations for valence and activation over a complete turn in order to obtain a ground truth for the turnwise AER experiments. Note that, unlike in databases annotated on the word level [15], short "activation peaks" like the stress of a single word within a sentence are unlikely to be captured by the annotators, due to the finite reaction time of the human labelers. Consequently, the time-continuous annotations tend to have low-pass characteristics and do not contain high frequencies, which limits the loss of information due to the averaging of annotation samples within a turn and accounts for the fact that emotion is perceived over a longer time window. The distribution of the averaged labels can be seen in Figs. 1 and 2. The dashed circle (with a radius of 0.33, dividing the axes into thirds) in the center of the valence-activation space in Fig. 1 marks a fifth region which represents a neutral emotional state. The coordinates that lie within this circle will be considered as belonging to a fifth, neutral class (see Section VI).

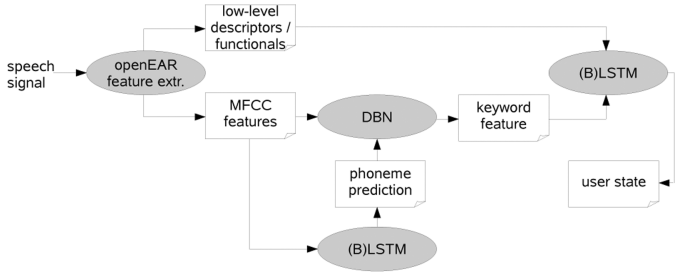

Fig. 3. Architecture of the acoustic-linguistic affect recognition system.

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 preselected *emotional prototypes* as in virtually any other database. Note that there is usually a high difference in accuracy between the tasks of prototypical and non-prototypical emotion recognition [23], [24], [46].

B. System Architecture

In Fig. 3, a flowchart of the presented incremental affect recognition system is shown. Processing components such as the LSTM network or the feature extractors are represented as ovals, whereas rectangles denote data. Depending on whether framewise or turnwise processing is used, our openEAR feature extractor module [47] (see Section IV) provides either low-level descriptors or statistical functionals of acoustic low-level features to the LSTM network (outlined in Section III) for emotion estimation. Additionally, mel-frequency cepstral coefficient (MFCC) features are provided to both components of the tandem keyword spotter component (see Section V), consisting of a DBN and a further LSTM network for phoneme prediction. Together with the produced phoneme predictions, the MFCC features are observed by the DBN, which then can detect the occurrence of a relevant keyword (i.e., a word that is relevant for valence or activation prediction, see Section V). Both, the discrete keyword feature and the acoustic features extracted by openEAR are used by an LSTM network to predict the user's current emotion. For the emotion coding, EmotionML[3] is used [48], [49], supporting continuous spatio-temporal emotion representation. EmotionML is a standard representation format for emotion-related states in technological contexts, developed by the W3C Emotion Markup Language Incubator Groups. It can be used within the tasks of data annotation, emotion recognition, and generation of emotion-related states.

Details about the overall architecture of the SEMAINE dialogue system can be found in [25].

Due to the complexity of the system, the computational cost of our AER engine is higher than for standard classification techniques such as SVMs, which however show significantly lower performance than the proposed system (see Section VI). Yet, when exclusively using *unidirectional* context within the LSTM framework, the causal system can operate in real-time: on an AMD Phenom 64 bit quad core CPU at 2.2 GHz, the openEAR feature extraction module runs online with a real-time factor (RTF) of 0.01, while the LSTM operates at a real-time factor of 0.09. Only one of the four cores was used for computation. Time and space complexity of the DBN is $\mathcal{O}(T \log T)$ and $\mathcal{O}(\log T)$, respectively, assuming that T corresponds to the length of the speech sequence that is currently processed.

[3]http://www.w3.org/2005/Incubator/emotion/XGR-emotionml-20081120/

III. Long Short-Term Memory

This section outlines the principle of the long short-term memory RNNs that are used for emotion classification in Section VI as well as for phoneme prediction in Section V. Framewise classification of emotion as investigated in this paper presumes a classifier that can access and model long-range context, since emotion mostly affects the long-term *dynamics* of prosodic, spectral, and voice quality speech features. When attempting to predict emotion frame by frame, a large number of preceding speech frames have to be taken into account in order to capture speech characteristics that are influenced by emotion. The *number* of speech frames which should be used to obtain enough context for reliably estimating emotion without affecting the capability of also detecting sudden changes of the speaker's emotional state is hard to determine [36], [37]. Thus, a classifier that is able to *learn* the amount of context is a promising alternative to manually defining fixed time windows for emotion recognition. Static techniques such as SVMs do not explicitly model context but rely on either capturing contextual information via statistical functionals of features [14] or aggregating frames using multi-instance learning techniques [50]. Dynamic classifiers like hidden Markov models are often used for flexible context modeling and time warping. Yet, HMMs have drawbacks such as the inherent assumption of conditional independence of successive observations, meaning that an observation is statistically independent of past observations provided that the values of the hidden variables are known. Hidden conditional random fields (HCRFs) [51] are one attempt to overcome this limitation. However, HCRF also offer no possibility to model a self-learned amount of contextual information. Other dynamic classifiers such as neural networks are able to model a certain amount of context by using cyclic connections. These so-called recurrent neural networks can in principle map from the entire *history* of previous inputs to each output. Yet, the analysis of the error flow in conventional recurrent neural nets led to the finding that long range context is inaccessible to standard RNNs since the backpropagated error either blows up or decays over time (vanishing gradient problem [52]). This led to the introduction of long short-term memory RNNs [27]. They are able to overcome the vanishing gradient problem and can learn the optimal amount of contextual information relevant for the classification task. Thus, LSTM architectures seem to be well-suited for our framewise emotion recognition task.

An LSTM layer is composed of recurrently connected memory blocks, each of which contains one or more memory cells, along with three multiplicative "gate" units: the input, output, and forget gates. The gates perform functions analogous to read, write, and reset operations. More specifically, the cell input is multiplied by the activation of the input gate, the cell output by that of the output gate, and the previous cell values by the forget gate (see Fig. 4). The overall effect is to allow the network to store and retrieve information over long periods of

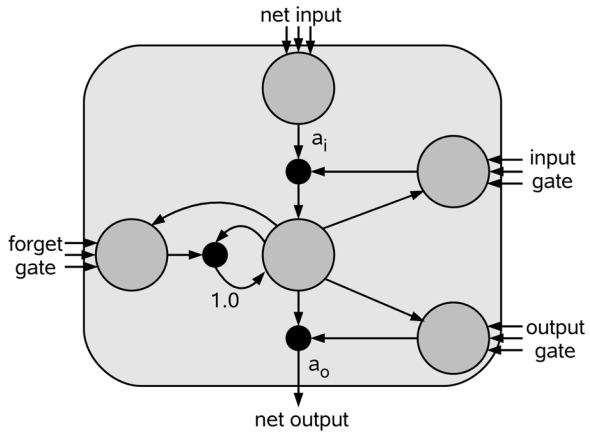

Fig. 4. LSTM memory block consisting of one memory cell: the input, output, and forget gates collect activations from inside and outside the block which control the cell through multiplicative units (depicted as small circles); input, output, and forget gate scale input, output, and internal state, respectively; a_i and a_o denote activation functions; the recurrent connection of fixed weight 1.0 maintains the internal state.

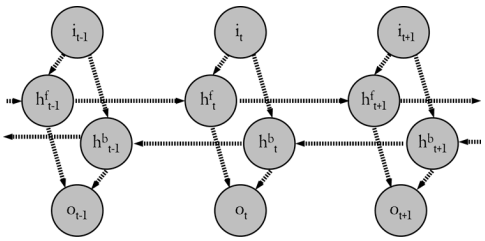

Fig. 5. Structure of a bidirectional network with input i, output o, and two hidden layers (h^f and h^b) for forward and backward processing.

TABLE I
28 LOW-LEVEL AUDIO FEATURES FOR TIME-CONTINUOUS EMOTION ANALYSIS (C) AND 39 FEATURES FOR TURN-BASED RECOGNITION (T); FEATURES IN BOLD FACE ARE USED FOR BOTH, CONTINUOUS AND TURN-BASED RECOGNITION

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 0, **MFCC 1-12**, MFCC 13-15	12	16
Time Signal	**Zero-Crossing-Rate**, max. and min. value, DC component	1	4
Spectral	**Energy in bands 0-250Hz, 0-650Hz, 250-650Hz, 1000-4000Hz**	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**

time. For example, as long as the input gate remains closed, the activation of the cell will not be overwritten by new inputs and can therefore be made available to the net much later in the sequence by opening the output gate.

Another problem with standard RNNs is that they have access to past but not to future context. This can be overcome by using bidirectional RNNs [40], where two separate recurrent hidden layers scan the input sequences in opposite directions. The two hidden layers are connected to the same output layer, which therefore has access to context information in both directions. The amount of context information that the network actually uses is learned during training, and does not have to be specified beforehand. Fig. 5 shows the structure of a simple bidirectional network.

Combining bidirectional networks with LSTM gives bidirectional LSTM [53], which has demonstrated excellent performance in phoneme recognition [28], [54], keyword spotting [29], and emotion recognition from speech [26].

While bidirectional LSTM cannot be used for online incremental prediction tasks, they are well suited to refine or correct the estimation of affect once the complete turn is available. Thus, we included bidirectional networks in our performance evaluation on the SAL database.

All RNN-based classifiers used in the experiments in Section VI were implemented using the open source RNNLIB library.[4]

TABLE II
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

IV. ACOUSTIC FEATURE EXTRACTION

Acoustic features from the speech signal are extracted using our openEAR [47] audio feature extractor, which was also used to provide features for the Interspeech 2009 Emotion Challenge [23].

The 28 low-level descriptors extracted from the audio signal for time-continuous emotion recognition are summarized in Table I (column "C'). The descriptors were extracted every 20 ms for overlapping frames with a frame-length of 32 ms. First-order regression coefficients are appended to the 28 low-level descriptors, resulting in a 56-dimensional feature vector for each frame.

In order to enable also turn-based emotion recognition experiments, the openEAR module alternatively follows the traditional approach of generating a large set of features by applying statistical functionals to low-level descriptor contours. An extended set of 39 low-level-descriptors detailed in Table I (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, resulting in a total of 4212 features. The 36 functionals are detailed in Table II.

The 4212 features for turn-based emotion recognition are reduced to relevant features for activation and valence independently by a correlation-based feature subset (CFS) selection

[4]http://github.com/alexgraves/RNNLIB

[55], [56]. The main idea of CFS is that useful feature subsets should contain features that are highly correlated with the target class while being uncorrelated with each other. The core of CFS is an evaluation function

$$M_S = \frac{k \cdot r_{\mathrm{cf}}}{\sqrt{k + k(k-1) r_{\mathrm{ff}}}} \quad (1)$$

where M_S is the rating of a subset S with k features. r_{cf} denotes the mean feature-class correlation and r_{ff} is the average feature-feature inter-correlation. Good subsets of features have highly predictive properties, yielding a high value in the numerator of (1), and a low degree of redundancy among the features, yielding a small value in the denominator. For correlation measurement, the symmetrical uncertainty coefficient is used (as described in [55]). To avoid an exhaustive search in the feature space a greedy hill climbing forward search is applied [56]. In this heuristic search algorithm, each feature is tentatively added to the feature subset, whereas the resulting set of features is evaluated using (1). Once the (so far) best feature set has been chosen, the procedure is repeated. Note that we willfully decided for a filter based feature selection method, since a wrapper-based technique would have biased the resulting feature set with respect to compatibility to a specific classifier.

Conducting CFS for turn-based emotion recognition via regression resulted in 60 features being selected for activation and 64 features for valence[5]. As termination criterion we considered a maximum of five non-improving nodes before terminating the greedy hill climbing forward search. Binary targets for activation and valence (high versus low, see Section VI) lead to the selection of 110 and 55 features, respectively. For the discriminative four-class quadrant classification task 121 features were selected, and for the five-class task applying CFS resulted in 123 selected features. Framewise emotion recognition uses the full set of $28 \cdot 2 = 56$ features without further reduction.

All features (turn-based functionals and low-level features) were standardized to have zero mean and unit standard deviation. These parameters were computed from the training data only and applied to both training and test data.

V. Linguistic Feature Extraction

This section outlines the tandem LSTM-DBN keyword spotter which generates binary linguistic features in order to incorporate knowledge about the spoken content via early fusion.

A. Background and References

Apart from acoustic features, also spoken or written text carries information about the underlying affective state [57]–[59]. This is usually reflected in the usage of certain words or grammatical alterations. A number of approaches exist for this analysis: keyword spotting [60], [61], rule-based modeling [62], semantic trees [63], Latent Semantic Analysis [64], transformation-based learning [65], world-knowledge-modeling [66], key-phrase spotting [67], and Bayesian networks [68], [69]. Two methods seem to be predominant, presumably because they are shallow representations of linguistic knowledge and have already been frequently employed in automatic speech processing: (class-based) N-grams [70]–[73] and vector space modeling [74], [75]. Due to the typical data sparseness in emotion recognition, unigrams mostly have been applied so far [72], [73]. The technique applied in our experiments is related to bag of words modeling [74]–[76] via keyword spotting; however, when applying framewise emotion recognition, only one keyword can be present at a given time frame. In the case of turnwise AER, the binary feature vector can contain more than one keyword. This would enable techniques like (bag of) N-gram modeling or other forms of linguistic information integration [77], [78], which however were not conducted in this paper in order to allow a fair comparison between framewise and turnwise affect recognition.

For combined acoustic and linguistic AER, the acoustic feature vector is extended by appending binary linguistic features. Each binary feature corresponds to the occurrence of one of the 56 keywords that were shown to be correlated to either valence or activation. 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."

When applying framewise acoustic-linguistic analysis, a short buffer 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 as linguistic features can still be delivered while the user is speaking. In order to reduce the vocabulary to a small set of emotionally meaningful keywords, correlation-based feature subset selection was applied on the training set. Pace regression [79]-based CFS used the continuous labels for valence and activation for bag of words keyword selection with a minimum term frequency of two (without stemming). Thereby 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*.[5]

The keyword spotter used in this paper is based on a recently introduced hierarchical DBN which was shown to significantly outperform a standard HMM-based approach [80]. The incorporation of an LSTM layer providing improved phoneme predictions was proven to further enhance keyword detection performance [41].

B. Design Overview

The tandem LSTM-DBN architecture we used for keyword spotting was proven to be robust with respect to phoneme recognition errors [41] and well suited for emotional speech. Its structure is depicted in Fig. 6. The network is composed of five different layers and hierarchy levels, respectively: a word layer, a phoneme layer, a state layer, the observed features, and the LSTM layer, consisting of inputs i_t, a hidden layer h_t, and outputs o_t (nodes inside the grey shaded box).

The following random variables are defined for every time step t: q_t denotes the phoneme identity, q_t^{ps} represents the position within the phoneme, q_t^{tr} indicates a phoneme transition,

[5]an explanation of the used features, openEAR configuration files, and lists of the selected features and keywords can be found at http://www.openaudio.eu/features_emo09.zip

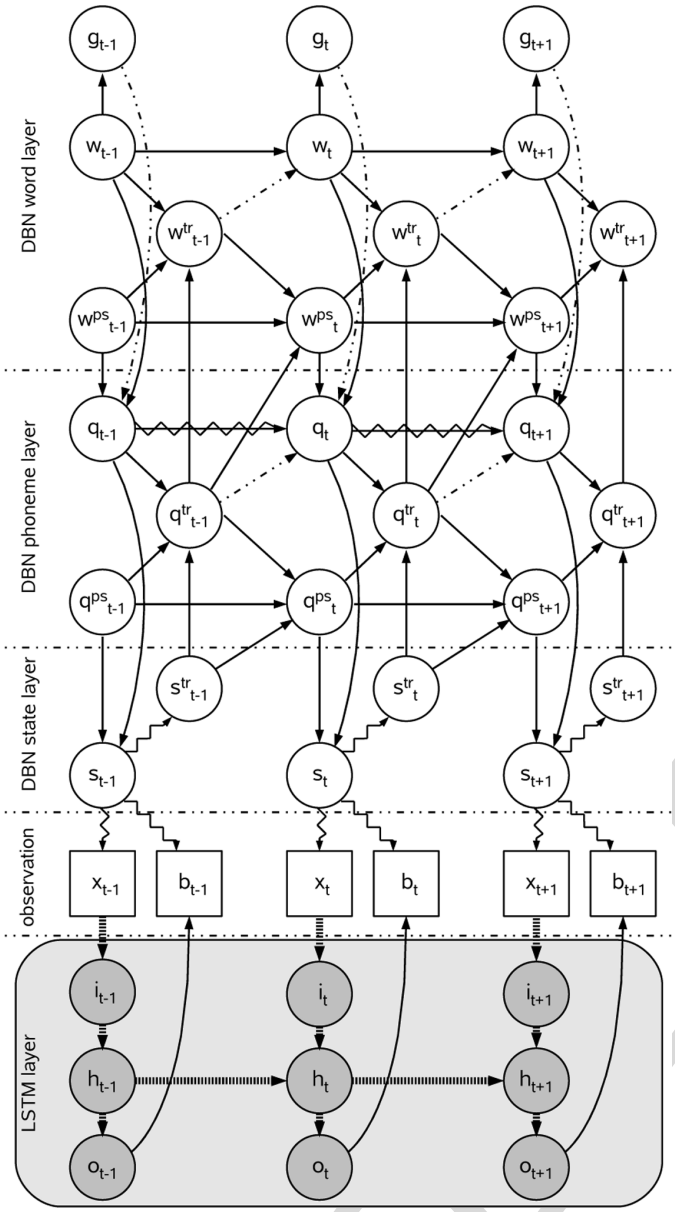

Fig. 6. Structure of the tandem LSTM-DBN keyword spotter: the LSTM network (gray shaded box) provides a discrete phoneme prediction feature b_t which is observed by the DBN, in addition to the MFCC features x_t. The DBN is composed of a state, phoneme, and word layer, consisting of hidden transition $(s_t^{\mathrm{tr}}, q_t^{\mathrm{tr}}, w_t^{\mathrm{tr}})$, position $(q_t^{\mathrm{ps}}, w_t^{\mathrm{ps}})$, and identity (s_t, q_t, w_t) variables. Hidden variables (circles) and observed variables (squares) are connected via random CPFs (zig-zagged lines) or deterministic CPFs (straight lines). Switching parent dependencies are indicated with dotted lines.

s_t is the current state with s_t^{tr} indicating a state transition, and x_t denotes the observed MFCC features. The variables w_t, w_t^{ps}, and w_t^{tr} are identity, position, and transition variables for the word layer of the DBN whereas a hidden *garbage variable* g_t indicates whether the current word is a keyword or not. A second observed variable b_t contains the phoneme prediction of the LSTM network. Fig. 6 displays hidden variables as circles and observed variables as squares. Deterministic conditional probability functions (CPFs) are represented by straight lines and zig-zagged lines correspond to random CPFs. Dotted lines refer to so-called *switching parents* [81], which allow a variable's parents to change conditioned on the current value of the switching parent. Note that a switching parent can not only change the set of parents but also the implementation (i.e., the CPF) of a parent. The bold dashed lines in the LSTM layer do not represent statistical relations but simple data streams.

C. Design Details

Assuming a speech sequence of length T, the DBN structure specifies the factorization

$$
\begin{aligned}
&p\left(g_{1:T}, w_{1:T}, w_{1:T}^{\mathrm{tr}}, w_{1:T}^{\mathrm{ps}}, q_{1:T}, q_{1:T}^{\mathrm{tr}}, q_{1:T}^{\mathrm{ps}}, s_{1:T}^{\mathrm{tr}}, s_{1:T}, x_{1:T}, b_{1:T}\right) \\
&\quad = \prod_{t=1}^{T} p(x_t \mid s_t) p(b_t \mid s_t) f\left(s_t \mid q_t^{\mathrm{ps}}, q_t\right) \\
&\qquad \times p\left(s_t^{\mathrm{tr}} \mid s_t\right) f\left(q_t^{\mathrm{tr}} \mid q_t^{\mathrm{ps}}, q_t, s_t^{\mathrm{tr}}\right) f(g_t \mid w_t) \\
&\qquad \times f\left(w_t^{\mathrm{tr}} \mid q_t^{\mathrm{tr}}, w_t^{\mathrm{ps}}, w_t\right) f\left(q_1^{\mathrm{ps}}\right) p\left(q_1 \mid w_1^{\mathrm{ps}}, w_1, g_1\right) \\
&\qquad \times f\left(w_1^{\mathrm{ps}}\right) p(w_1) \prod_{t=2}^{T} f\left(q_t^{\mathrm{ps}} \mid s_{t-1}^{\mathrm{tr}}, q_{t-1}^{\mathrm{ps}}, q_{t-1}^{\mathrm{tr}}\right) \\
&\qquad \times p\left(w_t \mid w_{t-1}^{\mathrm{tr}}, w_{t-1}\right) \\
&\qquad \times p\left(q_t \mid q_{t-1}^{\mathrm{tr}}, q_{t-1}, w_t^{\mathrm{ps}}, w_t, g_t\right) \\
&\qquad \times f\left(w_t^{\mathrm{ps}} \mid q_{t-1}^{\mathrm{tr}}, w_{t-1}^{\mathrm{ps}}, w_{t-1}^{\mathrm{tr}}\right)
\end{aligned} \tag{2}
$$

with $p(\cdot)$ denoting random conditional probability functions and $f(\cdot)$ describing deterministic CPFs.

The probability of the observed sequence can then be computed by summing over all hidden variables, whereas the factorization property in (2) can be exploited to optimally distribute the sums over the hidden variables into the products, using the junction tree algorithm [82].

The size of the LSTM input layer i_t corresponds to the dimensionality of the acoustic feature vector x_t, whereas the vector o_t contains one probability score for each of the P different phonemes at each time step. b_t is the index of the most likely phoneme:

$$
b_t = \max_{o_t}(o_{t,1}, \ldots, o_{t,j}, \ldots, o_{t,P}). \tag{3}
$$

The CPFs $p(x_t \mid s_t)$ are described by Gaussian mixtures, as is common practice with HMMs. Together with $p(b_t \mid s_t)$ and $p(s_t^{\mathrm{tr}} \mid s_t)$, they are learned via EM training. s_t^{tr} is a binary variable, indicating whether a state transition takes place or not. Since the current state is known with certainty, given the phoneme and the phoneme position, $f(s_t \mid q_t^{\mathrm{ps}}, q_t)$ is purely deterministic. A phoneme transition occurs whenever $s_t^{\mathrm{tr}} = 1$ and $q_t^{\mathrm{ps}} = S$ provided that S denotes the number of states of a phoneme. This is expressed by the function $f(q_t^{\mathrm{tr}} \mid q_t^{\mathrm{ps}}, q_t, s_t^{\mathrm{tr}})$. The phoneme position q_t^{ps} is known with certainty if s_{t-1}^{tr}, q_{t-1}^{ps}, and q_{t-1}^{tr} are given.

The hidden variable w_t can take values in the range $w_t = 0 \ldots K$ with K being the number of different keywords in the vocabulary. In case $w_t = 0$, the model is in the *garbage state* which means that no keyword is uttered at that time. The variable g_t is then equal to one. w_{t-1}^{tr} is a switching parent of w_t: if no word transition is indicated, w_t is equal to w_{t-1}. Otherwise, a word bigram specifies the CPF $p(w_t \mid w_{t-1}^{\mathrm{tr}} = 1, w_{t-1})$. In our experiments, we simplified the word bigram to a zerogram

which makes each keyword equally likely. However, we introduced differing *a priori* likelihoods for keywords and garbage phonemes:

$$p\left(w_t = 1 : K \mid w_{t-1}^{\mathrm{tr}} = 1\right) = \frac{K \cdot 10^a}{K \cdot 10^a + 1} \tag{4}$$

and

$$p\left(w_t = 0 \mid w_{t-1}^{\mathrm{tr}} = 1\right) = \frac{1}{K \cdot 10^a + 1}. \tag{5}$$

The parameter a can be used to adjust the tradeoff between true positives and false positives. Setting $a = 0$ means that the *a priori* probability of a keyword and the probability that the current phoneme does not belong to a keyword are equal. Adjusting $a > 0$ implies a more aggressive search for keywords, leading to higher true positive and false positive rates. The CPFs $f(w_t^{\mathrm{tr}} \mid q_t^{\mathrm{tr}}, w_t^{\mathrm{ps}}, w_t)$ and $f(w_t^{\mathrm{ps}} \mid q_{t-1}^{\mathrm{tr}}, w_{t-1}^{\mathrm{ps}}, w_{t-1}^{\mathrm{tr}})$ are similar to the phoneme layer of the DBN (i.e., the CPFs for q_t^{tr} and q_t^{ps}). However, we assume that "garbage words" always consist of only one phoneme, meaning that if $g_t = 1$, a word transition occurs as soon as $q_t^{\mathrm{tr}} = 1$. Consequently, w_t^{ps} is always zero if the model is in the garbage state. The variable q_t has two switching parents: q_{t-1}^{tr} and g_t. Similar to the word layer, q_t is equal to q_{t-1} if $q_{t-1}^{\mathrm{tr}} = 0$. Otherwise, the switching parent g_t determines the parents of q_t. In case $g_t = 0$—meaning that the current word is a keyword—q_t is a deterministic function of the current keyword w_t and the position within the keyword w_t^{ps}. If the model is in the garbage state, q_t only depends on q_{t-1} in a way that phoneme transitions between identical phonemes are forbidden.

Note that the design of the CPF $p(q_t \mid q_{t-1}^{\mathrm{tr}}, q_{t-1}, w_t^{\mathrm{ps}}, w_t, g_t)$ entails that the DBN will strongly tend to choose $g_t = 0$ (i.e., it will detect a keyword) once a phoneme sequence that corresponds to a keyword is observed. Decoding such an observation while being in the garbage state $g_t = 1$ would lead to "phoneme transition penalties' since the CPF $p(q_t \mid q_{t-1}^{\mathrm{tr}} = 1, q_{t-1}, w_t^{\mathrm{ps}}, w_t, g_t = 1)$ contains probabilities less than one. By contrast, $p(q_t \mid q_{t-1}^{\mathrm{tr}} = 1, w_t^{\mathrm{ps}}, w_t, g_t = 0)$ is deterministic, introducing no likelihood penalties at phoneme borders.

The DBN was implemented using the Graphical Models Toolkit (GMTK) [83]. In our experiments, we used phoneme models consisting of three states with 16 Gaussian mixtures. Phoneme models were trained on the TIMIT database [84] and adapted using the training split of the Sensitive Artificial Listener database (see Section II-A) to allow a better modeling of emotionally colored speech. Thereby all means, variances, and weights of the Gaussian mixture probability distributions $p(x_t \mid s_t)$, as well as the state transition probabilities $p(s_t^{\mathrm{tr}} \mid s_t)$ were re-estimated until the change of the overall log likelihood of the SAL training set became less than 0.02%. Since we found that in the context of our target application a low true positive rate is less critical than a high false positive rate, we chose a low tradeoff parameter of $a = 0$. The LSTM network of the tandem keyword spotter consists of 100 memory blocks of one cell each. All other DBN and LSTM parameters correspond exactly to those applied in [41]. Using these settings, the keyword spotter achieves a true positive rate of 0.59 at a false positive rate of 0.05 on the test partition of the SAL corpus.

VI. Experiments

Our emotion recognition engine was trained and tested on the SAL database (see Section II-A). In order to fit the requirements of the SEMAINE dialogue management [25], the recognition framework was designed in a way that it estimates the current quadrant in the two-dimensional valence-activation space. In addition to quadrant classification, we also investigated a five-class task including a "neutral" state, as well as discriminating low and high valence and activation separately.

A. Primary Systems Evaluated

For quadrant prediction we followed two different strategies: first, we trained LSTM networks for regression to obtain continuous predictions for valence and activation which were then mapped onto one of the four quadrants. In order to conduct feature selection independently for both the valence and the activation dimension, we used separate networks for the two dimensions. Second, the continuous labels for the emotional dimensions were mapped *before* training the network in order to allow a discriminative training on the quadrants, following the strategy introduced in [85]. These two strategies were also evaluated for the five-class task and for both of the two-class tasks (discrimination of low versus high activation and valence, respectively).

For each of the two techniques we evaluated both traditional turnwise classification with statistical functionals of acoustic features (see Section IV) and framewise classification using only low-level features. The gain of appending the binary keyword feature vector obtained by the dynamic Bayesian network (outlined in Section V) for combined acoustic-linguistic affect recognition was examined for every recognizer configuration.

The size of the LSTM input layer corresponds to the number of selected acoustic and linguistic features (see Sections IV and V), while the size of the output layer is equal to the number of regression/classification targets (one, two, four, and five, respectively). Each LSTM-RNN consists of one recurrent hidden layer with 50 memory blocks of one LSTM cell each. The BLSTM-RNN has two hidden layers of 50 memory blocks, one for each direction (forwards, backwards). For the acoustic–linguistic experiments the LSTM network size was increased to 70 memory blocks due to the increased size of the combined acoustic–linguistic feature vector. The networks were trained applying resilient propagation [86]. Prior to training, all weights were randomly initialized in the range from -0.1 to 0.1. Input and output gates used $\tanh$ activation functions, while the forget gates had logistic activation functions. Since the training converged faster for turnwise classification, we aborted turnwise training after ten epochs, whereas the training procedure for framewise classification was aborted after 250 epochs.

Before mapping the (B)LSTM-RNN predictions o_t onto quadrants, they were smoothed using a first-order low-pass filter to obtain the filtered predictions o_t^s

$$o_t^s = \alpha o_{t-1}^s + (1 - \alpha) \cdot o_t. \tag{6}$$

TABLE III
KAPPA VALUES FOR THE FOUR DIFFERENT ANNOTATORS IN THE SAL DATABASE (TURNWISE QUADRANT LABELING); ILA: INTER-LABELER AGREEMENT

κ	**1**	**2**	**3**	**4**
ILA	0.68	0.67	0.67	0.60
1		0.49	0.48	0.46
2			0.48	0.45
3				0.52

An α of 0.99 was used for time-continuous emotion recognition and an α of 0.7 was used for turn-based recognition. Both values were optimized on the training set.

B. Comparison Systems and Ground Truth

As a common continuous recognition technique, support vector regression (SVR) was performed for comparison [26], [56], [87]. The SVR used a polynomial kernel function of degree 1 and sequential minimal optimization (SMO). The discriminatively trained LSTM networks were compared to SVMs instead of SVR. Since SVR and SVM do not model contextual information, only turnwise classification was evaluated in this case. In order to determine the gain of long short-term memory modeling, we also investigated conventional RNN classification for comparison. The RNNs were trained in the same way as the LSTM networks; however, the network consisted of 50 hidden neurons instead of the 50 one-cell LSTM memory blocks.

Furthermore, we evaluated inter-labeler consistency as an upper benchmark for automatic emotion recognition. To obtain an impression of human emotion prediction quality we compared the annotations of one labeler to the mean of the annotations of the remaining three labelers. This was done for all of the four labelers so that eventually the average inter-labeler consistency could be determined.

As a further evaluation of inter-labeler agreement, Table III shows the kappa values for the four different annotators. Since each of the kappa values is larger than 0.4, the labeler agreement can be characterized as sufficiently high.

C. Results

Tables IV and VI show the recognition result for the assignment of quadrants using the regression method and the discriminative technique, respectively. Results for the five-class task which also considers a "neutral" state (see Fig. 1) can be seen in Tables V and VII, and Tables VIII and IX contain the results for separate classification of the degree of activation and valence (i.e., positive versus negative activation and valence, respectively). Due to the slightly unbalanced class distribution, accuracy is a rather inappropriate performance measure. Thus, we used the F1-measure as the harmonic mean between unweighted recall and unweighted precision for performance evaluation. Compared to emotion recognition on prototypical speech turns (as in [8] or [9]), the overall performance is significantly lower. Yet, the accuracies are in the order of magnitude that is typical for real-life experiments, attempting to classify natural, non-prototypical, and ambiguous emotional speech turns [23].

TABLE IV
REGRESSION-(B)LSTM AND RNN PERFORMANCE, SUPPORT VECTOR REGRESSION (SVR) PERFORMANCE, AND AVERAGE LABELER (LAB) CONSISTENCY FOR QUADRANT CLASSIFICATION USING TURNWISE OR FRAMEWISE PREDICTION WITH ACOUSTIC (A) OR ACOUSTIC–LINGUISTIC (A + L) FEATURES: ACCURACY (ACC.), UNWEIGHTED RECALL (REC.), UNWEIGHTED PRECISION (PREC.), AND F1-MEASURE (F1)

model	**unit**	**features**	**acc.**	**rec.**	**prec.**	**F1**
			quadrants			
BLSTM	turn	A	37.1 %	34.9 %	35.5 %	35.2 %
BLSTM	turn	A+L	41.0 %	36.9 %	37.8 %	37.3 %
BLSTM	frame	A	41.7 %	44.8 %	42.0 %	43.3 %
BLSTM	frame	A+L	48.2 %	51.6 %	49.3 %	**50.4 %**
LSTM	turn	A	37.3 %	37.9 %	35.4 %	36.6 %
LSTM	turn	A+L	38.6 %	38.4 %	39.8 %	39.7 %
LSTM	frame	A	31.2 %	33.4 %	37.2 %	35.2 %
LSTM	frame	A+L	34.2 %	30.7 %	37.9 %	33.9 %
RNN	turn	A	33.7 %	34.8 %	34.7 %	34.7 %
RNN	turn	A+L	37.1 %	35.5 %	36.7 %	36.1 %
RNN	frame	A	31.0 %	36.9 %	33.8 %	35.3 %
RNN	frame	A+L	28.2 %	31.7 %	34.8 %	33.2 %
SVR	turn	A	28.8 %	30.0 %	27.3 %	28.6 %
SVR	turn	A+L	33.3 %	32.2 %	30.4 %	31.3 %
lab	*turn*		*62.0 %*	*59.2 %*	*58.7 %*	*58.9 %*
lab	*frame*		*59.2 %*	*58.3 %*	*56.7 %*	*57.4 %*

TABLE V
REGRESSION-(B)LSTM AND RNN PERFORMANCE, SUPPORT VECTOR REGRESSION (SVR) PERFORMANCE, AND AVERAGE LABELER (LAB) CONSISTENCY FOR QUADRANT/NEUTRAL FIVE-CLASS TASK USING TURNWISE OR FRAMEWISE PREDICTION WITH ACOUSTIC (A) OR ACOUSTIC–LINGUISTIC (A + L) FEATURES: ACCURACY (ACC.), UNWEIGHTED RECALL (REC.), UNWEIGHTED PRECISION (PREC.), AND F1-MEASURE (F1)

model	**unit**	**features**	**acc.**	**rec.**	**prec.**	**F1**
			quadrants + neutral			
BLSTM	turn	A	37.9 %	34.1 %	38.6 %	36.2 %
BLSTM	turn	A+L	40.9 %	30.6 %	39.5 %	34.5 %
BLSTM	frame	A	34.6 %	39.3 %	34.3 %	36.6 %
BLSTM	frame	A+L	44.2 %	49.4 %	45.2 %	**47.2 %**
LSTM	turn	A	36.0 %	35.1 %	32.5 %	33.7 %
LSTM	turn	A+L	39.0 %	30.0 %	35.5 %	32.5 %
LSTM	frame	A	29.0 %	28.3 %	32.5 %	30.3 %
LSTM	frame	A+L	33.2 %	30.4 %	30.3 %	30.4 %
RNN	turn	A	35.1 %	30.9 %	33.2 %	32.0 %
RNN	turn	A+L	36.8 %	30.8 %	34.4 %	32.5 %
RNN	frame	A	35.6 %	21.1 %	41.4 %	27.9 %
RNN	frame	A+L	36.8 %	20.5 %	41.0 %	27.4 %
SVR	turn	A	32.8 %	25.5 %	24.9 %	25.2 %
SVR	turn	A+L	32.0 %	25.2 %	24.9 %	25.0 %
lab	*turn*		*56.8 %*	*55.1 %*	*53.7 %*	*54.3 %*
lab	*frame*		*56.3 %*	*56.9 %*	*54.9 %*	*55.8 %*

A rating of the prediction quality can be obtained when comparing the best result in Table IV (framewise BLSTM classification using acoustic and linguistic features) with the prediction performance of a human labeler (lab, frame in Table IV): when comparing the annotation of a single labeler to the mean of the annotations of the remaining three labelers, the obtained average F1-measure (57.4%) is only 7% higher than the F1-measure of the best classifier (50.4%). This reflects the ambiguity of perceived emotion and the resulting low degree of inter-labeler agreement. A further reason for the low annotator F1-measure is that a high amount of utterances are near the class borders (see Fig. 1). Consequently, those speech turns are hard to assign, even for human annotators. Such non-prototypical, ambiguous utterances also reduce the uncertainty during model training, which limits the obtainable automatic recognition performance.

TABLE VI
DISCRIMINATIVE (B)LSTM AND RNN PERFORMANCE, SUPPORT VECTOR MACHINE (SVM) PERFORMANCE, AND AVERAGE LABELER (LAB) CONSISTENCY FOR QUADRANT CLASSIFICATION USING TURNWISE OR FRAMEWISE PREDICTION WITH ACOUSTIC (A) OR ACOUSTIC–LINGUISTIC (A + L) FEATURES: ACCURACY (ACC.), UNWEIGHTED RECALL (REC.), UNWEIGHTED PRECISION (PREC.), AND F1-MEASURE (F1)

model	unit	features	acc.	rec.	prec.	F1
			quadrants			
BLSTM	turn	A	49.3 %	51.3 %	51.2 %	**51.3 %**
BLSTM	turn	A+L	47.6 %	48.6 %	46.8 %	47.7 %
BLSTM	frame	A	42.5 %	43.9 %	41.3 %	42.5 %
BLSTM	frame	A+L	39.0 %	37.4 %	37.1 %	37.2 %
LSTM	turn	A	48.6 %	47.4 %	48.2 %	47.8 %
LSTM	turn	A+L	44.9 %	49.1 %	48.3 %	48.7 %
LSTM	frame	A	37.4 %	38.0 %	38.1 %	38.1 %
LSTM	frame	A+L	32.0 %	37.8 %	32.6 %	35.3 %
RNN	turn	A	46.3 %	47.2 %	47.2 %	47.2 %
RNN	turn	A+L	45.9 %	46.5 %	45.8 %	46.1 %
RNN	frame	A	28.3 %	32.1 %	30.9 %	31.5 %
RNN	frame	A+L	22.1 %	28.2 %	27.3 %	27.7 %
SVM	turn	A	39.0 %	39.6 %	41.2 %	40.4 %
SVM	turn	A+L	37.8 %	38.5 %	36.7 %	37.6 %
lab	*turn*		*62.0 %*	*59.2 %*	*58.7 %*	*58.9 %*
lab	*frame*		*59.2 %*	*58.3 %*	*56.7 %*	*57.4 %*

TABLE VII
DISCRIMINATIVE (B)LSTM AND RNN PERFORMANCE, SUPPORT VECTOR REGRESSION (SVR) PERFORMANCE, AND AVERAGE LABELER (LAB) CONSISTENCY FOR QUADRANT/NEUTRAL FIVE-CLASS TASK USING TURNWISE OR FRAMEWISE PREDICTION WITH ACOUSTIC (A) OR ACOUSTIC–LINGUISTIC (A + L) FEATURES: ACCURACY (ACC.), UNWEIGHTED RECALL (REC.), UNWEIGHTED PRECISION (PREC.), AND F1-MEASURE (F1)

model	unit	features	acc.	rec.	prec.	F1
			quadrants + neutral			
BLSTM	turn	A	39.8 %	40.1 %	38.4 %	39.2 %
BLSTM	turn	A+L	41.9 %	41.8 %	41.7 %	**41.7 %**
BLSTM	frame	A	28.0 %	25.3 %	29.5 %	27.2 %
BLSTM	frame	A+L	29.0 %	32.3 %	25.8 %	28.7 %
LSTM	turn	A	40.0 %	38.7 %	36.0 %	37.3 %
LSTM	turn	A+L	41.9 %	41.5 %	37.1 %	39.2 %
LSTM	frame	A	27.8 %	28.6 %	29.6 %	29.1 %
LSTM	frame	A+L	30.4 %	30.0 %	24.7 %	27.1 %
RNN	turn	A	38.0 %	39.8 %	35.4 %	37.5 %
RNN	turn	A+L	39.0 %	41.6 %	37.1 %	39.2 %
RNN	frame	A	28.7 %	24.3 %	25.0 %	24.6 %
RNN	frame	A+L	27.0 %	25.6 %	26.4 %	26.0 %
SVM	turn	A	34.8 %	35.8 %	35.2 %	35.5 %
SVM	turn	A+L	34.8 %	35.9 %	35.0 %	35.4 %
lab	*turn*		*56.8 %*	*55.1 %*	*53.7 %*	*54.3 %*
lab	*frame*		*56.3 %*	*56.9 %*	*54.9 %*	*55.8 %*

TABLE VIII
REGRESSION-(B)LSTM AND RNN PERFORMANCE, SUPPORT VECTOR REGRESSION (SVR) PERFORMANCE, AND AVERAGE LABELER (LAB) CONSISTENCY FOR CLASSIFICATION OF VALENCE AND ACTIVATION (HIGH VERSUS LOW) USING TURNWISE OR FRAMEWISE PREDICTION WITH ACOUSTIC (A) OR ACOUSTIC–LINGUISTIC (A + L) FEATURES: ACCURACY (ACC.), UNWEIGHTED RECALL (REC.), UNWEIGHTED PRECISION (PREC.), AND F1-MEASURE (F1)

model	unit	features	acc.	rec.	prec.	F1
			activation			
BLSTM	turn	A	64.8 %	65.0 %	64.9 %	64.9 %
BLSTM	turn	A+L	64.1 %	64.3 %	64.1 %	64.2 %
BLSTM	frame	A	64.0 %	64.1 %	64.1 %	64.1 %
BLSTM	frame	A+L	65.7 %	65.7 %	65.6 %	**65.6 %**
LSTM	turn	A	59.8 %	60.9 %	61.3 %	61.1 %
LSTM	turn	A+L	60.2 %	60.7 %	60.7 %	60.7 %
LSTM	frame	A	56.4 %	57.2 %	57.4 %	57.3 %
LSTM	frame	A+L	59.1 %	59.9 %	60.1 %	60.0 %
RNN	turn	A	54.6 %	55.1 %	55.2 %	55.2 %
RNN	turn	A+L	55.6 %	56.4 %	56.5 %	56.5 %
RNN	frame	A	53.4 %	55.1 %	56.4 %	55.7 %
RNN	frame	A+L	49.3 %	49.4 %	49.4 %	49.4 %
SVR	turn	A	53.8 %	53.3 %	53.3 %	53.3 %
SVR	turn	A+L	55.5 %	55.2 %	55.8 %	55.2 %
lab	*turn*		*68.6 %*	*70.6 %*	*71.6 %*	*71.1 %*
lab	*frame*		*67.7 %*	*69.4 %*	*70.1 %*	*69.8 %*
			valence			
BLSTM	turn	A	56.5 %	58.0 %	58.3 %	58.1 %
BLSTM	turn	A+L	60.0 %	61.1 %	61.4 %	61.3 %
BLSTM	frame	A	65.8 %	64.0 %	64.7 %	64.3 %
BLSTM	frame	A+L	72.8 %	72.2 %	72.1 %	**72.2 %**
LSTM	turn	A	61.0 %	62.5 %	62.9 %	62.7 %
LSTM	turn	A+L	58.8 %	60.3 %	60.9 %	60.6 %
LSTM	frame	A	55.9 %	57.4 %	57.4 %	57.4 %
LSTM	frame	A+L	63.6 %	57.7 %	67.3 %	62.1 %
RNN	turn	A	58.8 %	60.3 %	60.8 %	60.5 %
RNN	turn	A+L	62.9 %	64.2 %	64.8 %	64.5 %
RNN	frame	A	60.9 %	63.6 %	64.3 %	63.9 %
RNN	frame	A+L	57.5 %	62.0 %	66.0 %	63.9 %
SVR	turn	A	53.1 %	55.0 %	55.6 %	55.3 %
SVR	turn	A+L	56.0 %	57.5 %	58.0 %	57.8 %
lab	*turn*		*88.6 %*	*88.4 %*	*88.6 %*	*88.6 %*
lab	*frame*		*86.0 %*	*85.8 %*	*85.6 %*	*85.7 %*

The best F1-measure for valence (72.2%) is notably below the average "performance" or consensus of a human labeler (85.7%). However, the best recognition result for activation (68.9%) is only 2.2% below the inter-human labeling consistency (71.1%). For the five-class task the performance gap between the best classifier and human labelers is 8.6% (see Table V).

In what follows, we will analyze the results in Tables IV–IX with respect to six different aspects: the number of emotion classes, the difference between regression and discriminative training, the gain of LSTM context modeling, the benefit of including bidirectional context, the difference between turnwise and framewise classification, and the integration of linguistic features.

1) Four Quadrants Versus Five Classes: The best F1-measure for quadrant classification can be obtained when using a discriminative BLSTM for turnwise prediction with acoustic features (51.3%, see Table VI). However, additionally modeling the "neutral" state can lead to a comparable prediction performance (47.2%, see Table V). Interestingly, for the five-class task framewise regression prevails. Obviously, the higher number of class borders a discriminative classifier has to face in the five-class experiment downgrades performance significantly. As can be seen in Table V, a BLSTM network modeling all five classes profits from frame by frame modeling of the fineness of emotional dynamics via regression. Tables X and XI show typical confusions when distinguishing four and five classes, respectively. In both cases, the best prediction quality can be obtained for quadrant IV (*angry/anxious*). Table XI points out that, due to the non-prototypicality of emotions in the SAL corpus, almost all quadrants are most frequently confused with the neutral state. An impression of the prediction quality for more prototypical utterances (or utterances with emotions of higher intensity) can be obtained when masking the last column and the last line of Table XI: quadrant–quadrant confusions obviously occur less frequent than quadrant–neutral confusions. Another interesting aspect is the effect of emotional intensity—and thus indirectly prototypicality—of the test set on the obtained recognition performance: when using the Regression-BLSTM for framewise prediction with acoustic and linguistic features (trained on *all*

TABLE IX
DISCRIMINATIVE-(B)LSTM AND RNN PERFORMANCE, SUPPORT VECTOR MACHINE (SVM) PERFORMANCE, AND AVERAGE LABELER (LAB) CONSISTENCY FOR CLASSIFICATION OF VALENCE AND ACTIVATION (HIGH VERSUS LOW) USING TURNWISE OR FRAMEWISE PREDICTION WITH ACOUSTIC (A) OR ACOUSTIC–LINGUISTIC (A + L) FEATURES: ACCURACY (ACC.), UNWEIGHTED RECALL (REC.), UNWEIGHTED PRECISION (PREC.), AND F1-MEASURE (F1)

model	unit	features	acc.	rec.	prec.	F1
			activation			
BLSTM	turn	A	68.3 %	68.9 %	68.8 %	**68.9 %**
BLSTM	turn	A+L	66.4 %	66.5 %	66.4 %	66.4 %
BLSTM	frame	A	62.8 %	63.6 %	64.0 %	63.8 %
BLSTM	frame	A+L	58.0 %	57.9 %	57.8 %	57.9 %
LSTM	turn	A	63.4 %	64.8 %	65.6 %	65.2 %
LSTM	turn	A+L	65.3 %	66.2 %	66.5 %	66.4 %
LSTM	frame	A	50.0 %	50.8 %	50.8 %	50.8 %
LSTM	frame	A+L	56.3 %	56.8 %	56.9 %	56.9 %
RNN	turn	A	61.7 %	63.0 %	63.8 %	63.4 %
RNN	turn	A+L	61.5 %	62.9 %	63.7 %	63.3 %
RNN	frame	A	50.6 %	52.7 %	53.8 %	53.3 %
RNN	frame	A+L	54.4 %	55.2 %	55.4 %	55.3 %
SVM	turn	A	55.8 %	56.7 %	56.8 %	56.8 %
SVM	turn	A+L	54.4 %	55.2 %	55.3 %	55.3 %
lab	*turn*		*68.6 %*	*70.6 %*	*71.6 %*	*71.1 %*
lab	*frame*		*67.7 %*	*69.4 %*	*70.1 %*	*69.8 %*
			valence			
BLSTM	turn	A	63.7 %	64.6 %	64.7 %	64.7 %
BLSTM	turn	A+L	71.2 %	71.8 %	71.7 %	**71.7 %**
BLSTM	frame	A	63.8 %	65.1 %	64.8 %	65.0 %
BLSTM	frame	A+L	55.0 %	58.4 %	59.7 %	59.0 %
LSTM	turn	A	56.4 %	59.4 %	63.4 %	61.3 %
LSTM	turn	A+L	66.8 %	68.5 %	70.1 %	69.3 %
LSTM	frame	A	65.3 %	66.3 %	65.9 %	66.1 %
LSTM	frame	A+L	58.3 %	56.1 %	56.6 %	56.4 %
RNN	turn	A	67.5 %	67.9 %	67.8 %	67.9 %
RNN	turn	A+L	69.5 %	70.5 %	70.6 %	70.5 %
RNN	frame	A	57.5 %	60.3 %	61.0 %	60.6 %
RNN	frame	A+L	64.2 %	64.6 %	64.2 %	64.4 %
SVM	turn	A	61.4 %	63.5 %	65.7 %	64.6 %
SVM	turn	A+L	59.3 %	61.4 %	62.9 %	62.1 %
lab	*turn*		*88.6 %*	*88.4 %*	*88.6 %*	*88.6 %*
lab	*frame*		*86.0 %*	*85.8 %*	*85.6 %*	*85.7 %*

TABLE X
CONFUSION MATRIX FOR THE BEST QUADRANT CLASSIFICATION SETTING (DISCRIMINATIVE BLSTM FOR TURNWISE PREDICTION WITH ACOUSTIC FEATURES ONLY); ROWS: GROUND TRUTH; COLUMNS: PREDICTIONS (WHITE TO BLACK RESEMBLES 0–100 %)

%	I	II	III	IV
I	39	31	9	21
II	9	54	12	25
III	4	27	47	22
IV	3	21	9	67

training data and characterized by the five-class confusion matrix in Table XI), while evaluating only those utterances that are *not* annotated as "neutral," the resulting quadrant prediction F1-measure is 58.2%. On the other hand, when evaluating only those turns that are annotated as "neutral," the F1-measure for quadrant prediction is as low as 34.3%. For very "intense" test utterances that are labeled as having an absolute value of activation and valence that is higher than 0.5, the obtained quadrant prediction F1-measure is 85.1%.

2) Regression Versus Discriminative Training: For almost every experimental setting we can observe that discriminative

TABLE XI
CONFUSION MATRIX FOR THE BEST "QUADRANTS + NEUTRAL" (N) CLASSIFICATION SETTING (REGRESSION BLSTM FOR FRAMEWISE PREDICTION WITH ACOUSTIC AND LINGUISTIC FEATURES); ROWS: GROUND TRUTH; COLUMNS: PREDICTIONS (WHITE TO BLACK RESEMBLES 0–100 %)

%	I	II	III	IV	N
I	40	13	6	4	37
II	25	40	3	8	24
III	12	1	48	14	25
IV	2	9	1	80	8
N	22	11	10	16	41

training prevails for turnwise recognition while regression prevails for framewise recognition. Complete turns that are characterized by statistical functionals of features can be distinguished better with a discriminative technique. On the other hand, when predicting a class frame by frame the network fails to model "label jumps" when discriminatively trained on the discrete labels. For framewise prediction, modeling the smooth progression of valence and activation is necessary before mapping the output activations to quadrants.

3) LSTM Context Modeling Versus RNN and SVM: Both, for framewise but also for turnwise prediction the LSTM architecture outperforms a conventional RNN in most cases. The major reason for this is the *vanishing gradient problem* (see Section III) which limits the amount of context a recurrent neural network can access. Using no contextual information at all leads to comparatively low performance as can be seen in the SVR and SVM experiments, justifying the higher computational cost of the LSTM approach.

4) Unidirectional Versus Bidirectional Context: Independent of the classification task, bidirectional context mostly prevails over unidirectional context. Both, regression and discriminative BLSTM networks outperform all other models (LSTM, RNN, SVR, and SVM) for the discrimination of five, four, and two classes (numbers in bold face in Tables IV–IX).

5) Turnwise Versus Framewise Classification: As already mentioned, turnwise prediction can successfully be combined with discriminative learning, while framewise emotion recognition is rather suited for predictors based on regression. For both strategies, modeling contextual information is essential. When additionally modeling "neutrality," the best result can be obtained with framewise prediction (see Table V). Note that the amount of contextual information a BLSTM network models is a lot more flexible when framewise prediction is applied, since the temporal granularity is higher than it is for turnwise recognition. This can be seen as the major reason why framewise recognition outperforms turnwise prediction if regression-BLSTM networks are used.

6) Acoustic Features Versus Combined Acoustic and Linguistic Features: Comparing Tables IV and VI, one can assert that the regression-LSTM seems to profit more from the inclusion of linguistic features. In some cases the quadrant prediction performance of the discriminative classifier is even degraded when adding keyword features. Obviously, the presence of single keywords is not discriminative enough in this case. Linguistic features are rather suited for modeling tendencies within a continuous scale for valence and activation. When modeling

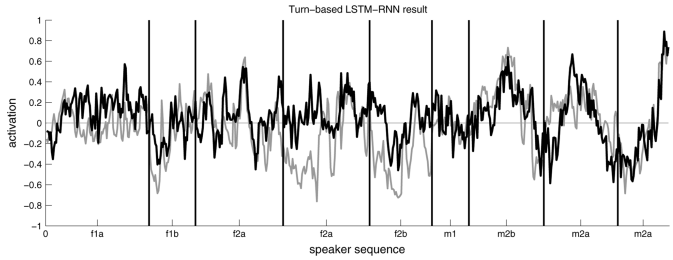

Fig. 7. Prediction of activation (black) using a regression-LSTM and ground truth (gray) over all turns of the test set (only acoustic features used).

"neutrality" as a fifth class, also the *discriminative* BLSTM profits from linguistic features (while this is not the case for the discriminative *four-class* task). This supports the finding that a performance gain through keyword features presumes a certain level of granularity of the prediction targets.

As an example for emotion recognition using regression, Fig. 7 shows the turnwise activation predictions of a regression-LSTM before the output activations are mapped onto quadrants. Prediction and ground truth are correlated with a correlation coefficient of 0.56, leading to an F1-measure of 61.1% (see Table VIII) when distinguishing positive and negative activation for every speech turn.

VII. Conclusion

In this paper, we introduced a novel technique for the estimation of the quadrant in a two-dimensional emotional space spanned by the dimensions *valence* and *activation*, as it is needed for the SAL—an emotionally sensitive virtual agent developed within the SEMAINE project. In contrast to many other works that report recognition results for the static classification of acted speech turns representing *emotional prototypes*, our contribution can be seen as a realistic evaluation of recognition accuracy under real-life conditions, where non-prototypical speech has to be classified using powerful techniques of dynamic speech modeling. Our approach combines acoustic features obtained by our openEAR online feature extractor with binary linguistic features produced by a tandem LSTM-DBN, which are then classified by a long short-term memory recurrent neural net. The LSTM architecture allows for the modeling of long-range contextual information and enables a new technique of incremental affect recognition that does not require the computation of statistical functionals of features but captures the temporal evolution indirectly through LSTM memory cells. As an alternative for regression-based quadrant prediction, we designed a discriminatively trained LSTM network which explicitly learns to distinguish quadrants of the emotional space. The design of our proposed AER system is based on a series of findings documented in earlier works: the benefit of including linguistic features for speech based emotion recognition [14], the enhancement of keyword spotting performance through the incorporation of LSTM phoneme prediction features [41], the importance of modeling temporal long-range dependencies in emotion recognition [26], and the potential of discriminative learning for quadrant prediction [85]. The prediction quality of our system was shown to be comparable to the degree of consistency between different human labelers.

One short-coming of our system is the fact that *bidirectional* context cannot be used in a causal online emotion recognition system. However, since we observed improved results for bidirectional LSTM networks, the investigation of the potential of BLSTM-RNN for online recognition is promising. For future experiments, 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. A further drawback of the introduced system is its complexity. However, provided that only *unidirectional* context is used, our system can still operate in real-time. The training of the complete system as used in this paper can be completed within one day, but will take longer as soon as larger training databases are used. Another problem—implied by the recognition task—is that our classification system has to deal with a high amount of ambiguous speech turns which are near the class borders in the valence-activation space. This leads to high error rates for non-prototypical speech segments that are difficult to model when using discrete classes. A possible solution is to continuously model emotion via regression while abstaining from mapping the regression output onto quadrants. Yet, those continuous values are difficult to use for the dialogue management system of an emotion-sensitive virtual agent which will have to use thresholds or any other kind of discretization before selecting adequate system responses. As far as AER performance evaluation is concerned, a possible solution is to increase the granularity of emotional space discretization (e.g., by defining nine instead of four regions in the emotional space) while at the same time tolerating confusions between neighboring regions, as done in [26], for example. Even though "wrong" assignments of ambiguous speech turns are not necessarily critical for the quality or adequateness of a virtual agent's responses (even humans can interpret such utterances differently), further research will be necessary in this area.

Future works will focus on investigating the benefit of including further feature types, such as vision features used in [14] or [88], into a time-continuous context sensitive emotion recognition framework. For this purpose it would be interesting to examine the potential of hybrid fusion techniques such as asynchronous hidden Markov models [89] or multidimensional dynamic time warping [90] as alternatives to late and early fusion. Also the LSTM architecture and parameterization could be optimized by including more hidden layers or using different layer sizes. Furthermore it would be interesting to examine the potential of multi-task learning, i.e., learning the phonemes and the affective state simultaneously. In addition to the mentioned approaches for future improvements, there will be a lot more aspects to consider before emotion-sensitive systems can show a degree of naturalness that is comparable to humans. Yet, 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, the principle of incremental speech processing and the integration of long-range context information can be seen as two further steps towards making virtual agents more human-like.

REFERENCES

[1] M. T. Vo and A. Waibel, "Multimodal human–computer interaction," in *Proc. ISSD'93*, Waseda, Japan, 1993.

[2] S. Furui, "Toward the ultimate synthesis/recognition," *Proc. Nat. Acad. Sci. USA*, vol. 92, no. 22, pp. 10 040–10 045, 1995.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human–computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Feb. 2001.

[4] R. W. Picard, "Toward agents that recognize emotion," in *Actes Proc. IMAGINA*, 1998, pp. 153–165.

[5] E. Shriberg, "How peoply really talk and why engineers should care," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 1781–1784.

[6] Z. Zeng, M. Pantic, G. I. Rosiman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[7] R. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.

[8] S. Casale, A. Russo, G. Scebba, and S. Serrano, "Speech emotion classification using machine learning algorithms," in *Proc. IEEE Int. Conf. Semantic Comput.*, 2008, pp. 158–165.

[9] B. Schuller, M. Wimmer, L. Mösenlechner, C. Kern, D. Arsic, and G. Rigoll, "Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space?," in *Proc. ICASSP'08*, Las Vegas, NV, 2008, pp. 4501–4504.

[10] M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[11] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Netw.*, vol. 18, no. 4, pp. 407–422, 2005.

[12] ***[AUTHOR: Please provide page range]***V. Petrushin, "Emotion in speech: Recognition and application to call centers," *Artif. Neural Netw Eng. (ANNIE)*, 1999.

[13] B. Schuller, R. Müller, B. Hörnler, A. Hoethker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *Proc. Int. Conf. Multimodal Interfaces, ACM SIGHI*, Nagoya, Japan, 2007, pp. 30–37.

[14] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? recognising natural interest by extensive audiovisual integration for real-life application," *Image Vis. Comput. J. (IMAVIS), Special Iss. Vis. Multimodal Anal. Human Spontaneous Behavior*, vol. 27, no. 12, pp. 1760–1774, 2009.

[15] A. Batliner, S. Steidl, and E. Nöth, "Releasing a thoroughly annotated and processed spontaneous emotional database: The FAU Aibo Emotion Corpus," in *Proc. Satellite Workshop of LREC 2008 Corpora Res. Emotion and Affect*, L. Devillers, J. C. Martin, R. Cowie, E. Douglas-Cowie, and A. Batliner, Eds., 2008, pp. 28–31.

[16] S. Steininger, F. Schiel, O. Dioubina, and S. Raubold, "Development of user-state conventions for the multimodal corpus in smartkom," in *Proc. Workshop Multimodal Resources Multimodal Syst. Eval.*, Las Palmas, Spain, 2002, pp. 33–37.

[17] B. Schuller, G. Rigoll, S. Can, and H. Feussner, "Emotion sensitive speech control for human-robot interaction in minimal invasive surgery," in *Proc. 17th Int. Symp. Robot Human Interactive Commun. RO-MAN'08*, Munich, Germany, 2008, pp. 453–458.

[18] J. Hansen and S. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 1743–1746.

[19] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Proc. ICME*, Amsterdam, The Netherlands, 2005.

[20] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech'05*, Lisbon, Portugal, 2005, pp. 1517–1520.

[21] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a Danish emotional speech database," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 1695–1698.

[22] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 Audiovisual Emotion Database," in *Proc. IEEE Workshop Multimedia Database Management*, Atlanta, Georgia, 2006.

[23] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 emotion challenge," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 312–315.

[24] S. Steidl, B. Schuller, A. Batliner, and D. Seppi, "The hinterland of emotions: Facing the open-microphone challenge," in *Proc. ACII*, Amsterdam, The Netherlands, 2009, pp. 690–697.

[25] M. Schröder, R. Cowie, D. Heylen, M. Pantic, C. Pelachaud, and B. Schuller, "Towards responsive sensitive artificial listeners," in *Proc. 4th Int. Workshop Human-Comput. Convers.*, Bellagio, Italy, 2008.

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

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, no. 5–6, pp. 602–610, Jun. 2005.

[29] M. Wöllmer, F. Eyben, J. Keshet, A. Graves, B. Schuller, and G. Rigoll, "Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks," in *Proc. ICASSP*, Taipei, Taiwan, 2009, pp. 3949–3952.

[30] M. Wöllmer, F. Eyben, B. Schuller, Y. Sun, T. Moosmayr, and N. Nguyen-Thien, "Robust in-car spelling recognition—A tandem BLSTM-HMM approach," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 2507–2510.

[31] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural Comput. Applicat.*, vol. 9, pp. 290–296, 2000.

[32] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[33] B. Schuller, S. Reiter, and G. Rigoll, "Evolutionary feature generation in speech emotion recognition," in *Proc. ICME'06*, Toronto, ON, Canada, 2006, pp. 5–8.

[34] M. Streit, A. Batliner, and T. Portele, "Emotions analysis and emotion-handling subdialogues," in *SmartKom: Foundations of Multimodal Dialogue Systems*, W. Wahlster, Ed. Berlin, Germany: Springer, 2006, pp. 317–332.

[35] S. Steidl, in *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*, Berlin, Germany, 2009, Logos Verlag.

[36] B. Schuller and G. Rigoll, "Timing levels in segment-based speech emotion recognition," in *Proc. Interspeech'06*, Pittsburgh, PA, 2006, pp. 1818–1821, ISCA.

[37] B. Schuller, B. Vlasenko, R. Minguez, G. Rigoll, and A. Wendemuth, "Comparing one and two-stage acoustic modeling in the recognition of emotion in speech," in *Proc. ASRU'07*, Kyoto, Japan, 2007, pp. 596–600.

[38] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. ICASSP'03*, Hong Kong, China, 2003, pp. 1–4.

[39] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. turnlevel: Emotion recognition from speech considering static and dynamic processing," in *Proc. ACII'07, Lisbon, Portugal*, A. Paiva, Ed., Heidelberg, Germany, 2007, vol. LNCS 4738, pp. 139–147, Springer Berlin.

[40] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[41] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll, "A tandem BLSTM-DBN architecture for keyword spotting with enhanced context modeling," in *Proc. NOLISP*, Vic, Spain, 2009.

[42] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J. C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data," in *Affective Computing and Intelligent Interaction*. Berlin/Heidelberg, Germany: Springer, 2007, vol. 4738/2007, pp. 488–500 [Online]. Available: http://emotionresearch.net/download/pilot-db/

[43] C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 1983–1986.

[44] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "FEELtrace: An instrument for recording perceived emotion in real time," in *Proc. ISCA Workshop Speech Emotion*, 2000, pp. 19–24.

[45] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME)*, Hannover, Germany, 2008, pp. 865–868.

[46] D. Seppi, A. Batliner, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and V. Aharonson, "Patterns, prototypes, performance: Classifying emotional user states," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 601–604.
[47] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR—Introducing the Munich open-source emotion and affect recognition toolkit," in *Proc. ACII*, Amsterdam, The Netherlands, 2009, pp. 576–581.
[48] M. Schröder, L. Devillers, K. Karpouzis, J.-C. Martin, C. Pelachaud, C. Peter, H. Pirker, B. Schuller, J. Tao, and I. Wilson, "What should a generic emotion markup language be able to represent?," in *Affective Computing and Intelligent Interaction*, A. Paiva, R. Prada, and R. W. Picard, Eds. Berlin-Heidelberg, Germany: Springer, 2007, pp. 440–451.
[49] P. Baggia, F. Burkhardt, C. Pelachaud, C. Peter, B. Schuller, I. Wilson, and E. Zovato, "Elements of an EmotionML 1.0," in *W3C Incubator Group Report*, M. Schröder, Ed., 2008, W3C.
[50] B. Schuller and G. Rigoll, "Recognising interest in conversational speech-comparing bag of frames and supra-segmental features," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 1999–2002.
[51] A. Quattoni, S. Wang, L. P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1848–1853, Oct. 2007.
[52] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "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. Piscataway, NJ: IEEE Press, 2001.
[53] A. Graves, S. Fernandez, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Proc. ICANN*, Warsaw, Poland, 2005, vol. 18, pp. 602–610.
[54] S. Fernandez, A. Graves, and J. Schmidhuber, "Phoneme Recognition in TIMIT With BLSTM-CTC," IDSIA, Tech. Rep., 2008.
[55] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Univ. of Waikato, Hamilton, New Zealand, 1999.
[56] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.
[57] S. Arunachalam, D. Gould, E. Anderson, D. Byrd, and S. Narayanan, "Politeness and frustration language in child-machine interactions," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 2675–2678.
[58] Z. J. Chuang and C. H. Wu, "Emotion recognition using acoustic features and textual content," in *Proc. ICME*, 2004, pp. 53–56.
[59] K. Dupuis and K. Pichora-Fuller, "Use of lexical and affective prosodic cues to emotion by younger and older adults," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 2237–2240.
[60] C. Elliott, "The affective reasoner: A process model of emotions in a multi-agent system," Ph.D. dissertation, Northwestern Univ., Evanston, IL, 1992.
[61] R. Cowie, E. Douglas-Cowie, B. Apolloni, J. Taylor, A. Romano, and W. Fellenz, "What a neural net needs to know about emotion words," *Comput. Intell. Applicat.*, pp. 109–114, 1999, N. Mastorakis, Ed..
[62] D. Litman and K. Forbes, "Recognizing emotions from student speech in tutoring dialogues," in *Proc. ASRU*, Virgin Islands, 2003, pp. 25–30.
[63] X. Zhe and A. Boucouvalas, "Text-to-emotion engine for real time internet communication," in *Proc. Int. Symp. Commun. Syst., Netw., DSPs*, 2002, pp. 164–168, Staffordshire Univ., Stoke-on-Trent, U.K.
[64] B. Goertzel, K. Silverman, C. Hartley, S. Bugaj, and M. Ross, "The baby webmind project," in *Proc. Annu. Conf. Soc. Study Artif. Intell. Simul. Behav. (AISB)*, 2000.
[65] T. Wu, F. Khan, T. Fisher, L. Shuler, and W. Pottenger, "Posting act tagging using transformation-based learning," in *Foundations of Data Mining and Knowledge Discovery*, T. Y. Lin, S. Ohsuga, C. J. Liau, X. Hu, and S. Tsumoto, Eds., 2005, pp. 319–331.
[66] H. Liu, H. Liebermann, and T. Selker, "A model of textual affect sensing using real-world knowledge," in *Proc. IUI*, 2003, pp. 125–132.
[67] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Proc. ICASSP*, 2004, pp. 577–580.
[68] J. Breese and G. Ball, "Modeling emotional state and personality for conversational agents," Microsoft, Tech. Rep., 1998.
[69] G. Rigoll, R. Müller, and B. Schuller, "Speech emotion recognition exploiting acoustic and linguistic information sources," in *Proc. SPECOM*, Patras, Greece, 2005, pp. 61–67.
[70] T. S. Polzin and A. Waibel, "Emotion-sensitive human–computer interfaces," in *Proc. ISCA ITRW Speech Emotion*, 2000, pp. 201–206.
[71] J. Ang, R. Dhillon, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human–computer dialog," in *Proc. Interspeech*, Denver, CO, 2002, pp. 2037–2040.
[72] C. M. Lee, S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition," in *Proc. ICSLP*, 2002, pp. 873–376.
[73] L. Devillers, L. Lamel, and I. Vasilescu, "Emotion detection in task-oriented spoken dialogs," in *Proc. ICME*, Baltimore, MD, 2003.
[74] B. Schuller, R. Müller, M. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensemble," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 805–808.
[75] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "Combining efforts for improving automatic classification of emotional user states," in *Proc. IS-LTC*, Ljubliana, Slovenia, 2006, pp. 240–245.
[76] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. ECML*, Chemniz, Germany, 1998, pp. 137–142.
[77] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Emotion recognition from speech: Putting asr in the loop," in *Proc. ICASSP*, Taipei, Taiwan, 2009, pp. 4585–4588.
[78] B. Schuller, J. Schenk, and G. Rigoll, ""The Godfather" vs. "chaos": Comparing linguistic analysis based on online knowledge sources and bags-of-n-grams for movie review valence estimation," in *Proc. ICDAR*, Barcelona, Spain, 2009, pp. 858–862.
[79] Y. Wang and I. H. Witten, "Modeling for optimal probability prediction," in *Proc. 19th Int. Conf. Mach. Learn.*, Sydney, Australia, 2002, pp. 650–657.
[80] M. Wöllmer, F. Eyben, B. Schuller, and G. Rigoll, "Robust vocabulary independent keyword spotting with graphical models," in *Proc. ASRU*, Merano, Italy, 2009, pp. 349–353.
[81] J. A. Bilmes, "Graphical models and automatic speech recognition," in *Mathematical Foundations of Speech and Language Processing*, R. Rosenfeld, M. Ostendorf, S. Khudanpur, and M. Johnson, Eds. New York: Springer Verlag, 2003, pp. 191–246.
[82] F. V. Jensen, *An Introduction to Bayesian Networks*. New York: Springer, 1996.
[83] J. Bilmes and G. Zweig, "The graphical models toolkit: An open source software system for speech and time-series processing," in *Proc. ICASSP*, 2002, pp. 3916–3919.
[84] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," 1993.
[85] M. Wöllmer, F. Eyben, B. Schuller, E. Douglas-Cowie, and R. Cowie, "Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 1595–1598.
[86] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in *Proc. IEEE Int. Conf. Neural Netw.*, 1993, pp. 586–591.
[87] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *Proc. ICASSP*, Honolulu, HI, 2007, pp. 1085–1088.
[88] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: Face, body gesture, speech," in *Affect and Emotion in Human-Computer Interaction*, C. Peter and R. Beale, Eds. New York: Springer, 2008, vol. 4868, lNCS.
[89] S. Bengio, "An asynchronous hidden Markov model for audio-visual speech recognition," Advances in NIPS 15 2003.
[90] M. Wöllmer, M. Al-Hames, F. Eyben, B. Schuller, and G. Rigoll, "A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams," *Neurocomputing*, vol. 73, pp. 366–380, 2009.

Martin Wöllmer (M'09) received the diploma in electrical engineering and information technology from the Technische Universität München (TUM), Munich, Germany.

He works as a Researcher funded by the European Community's Seventh Framework Program project SEMAINE (FP7/2007–2013) at TUM, where his current research and teaching activity includes the subject areas of pattern recognition and speech processing. His focus lies on multimodal data fusion, automatic recognition of emotionally colored and noisy speech, and speech feature enhancement. Publications of his in various

journals and conference proceedings cover novel and robust modeling architectures for speech and emotion recognition such as switching linear dynamic models or long short-term memory recurrent neural nets.

Mr. Wöllmer is a Reviewer for the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING and the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING.

Björn Schuller (M'04) received the diploma and Ph.D. degrees in electrical engineering and information technology from Technische Universität München (TUM), Munich, Germany.

He is currently a Lecturer in pattern recognition at TUM. He authored more than 120 publications in books, journals, and peer-reviewed conference proceedings in this field. Best known are his works advancing audiovisual processing in the areas of affective computing and multimedia retrieval.

Dr. Schuller serves as a member of the steering committee of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, as associate editor and reviewer for several scientific journals, including the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, *Elsevier Signal Processing*, and *Speech Communication*, and as invited speaker, session organizer, and chairman, and program committee member of numerous international conferences. Current project steering board activities include SEMAINE funded by the European Community and further projects with companies as BMW, Continental, Daimler, Siemens, Toyota, and VDO. He is invited expert in the W3C Emotion and Emotion Markup Language Incubator Groups, and elected member of HUMAINE Association Executive Committee.

Florian Eyben (M'09) received the diploma in information technology from the Technische Universität München (TUM), Munich, Germany.

He works on a research grant as part of the European Community's Seventh Framework Program project SEMAINE (FP7/2007–2013)—The Sensitive Artificial Listener project—within the Institute for Human–Machine Communication at TUM. Teaching activities of his comprise pattern recognition and speech and language processing. His research interests include large-scale hierarchical audio feature extraction and evaluation, automatic emotion recognition from the speech signal, recognition of non-linguistic vocalizations, automatic continuous large-vocabulary speech recognition, statistical and context-dependent language models, and music information retrieval. He has several publications in various journals and conference proceedings covering many of his areas of research.

Gerhard Rigoll (M'86–SM'98) received the diploma in technical cybernetics in 1982, the Ph.D. degree for his work in the field of automatic speech recognition in 1986, and the habilitation in the field of speech synthesis in 1991, all from the University of Stuttgart, Stuttgart, Germany.

He was with the Fraunhofer-Institute Stuttgart, Speech Plus in Mountain View, CA, and Digital Equipment in Maynard, MA, spent a Post-Doctoral Fellowship at IBM T. J. Watson Research Center, Yorktown Heights, NY, headed a research group at Fraunhofer-Institute Stuttgart, and spent a two year's research stay at NTT Human Interface Laboratories in Tokyo, Japan, in 1986, in the area of neuro-computing, speech recognition and pattern recognition until he was appointed as a Full Professor of computer science at Gerhard-Mercator-University, Duisburg, Germany, 1993 and of Human–Machine Communication at the Technische Universität München (TUM), Munich, Germany, in 2002. He authored and coauthored more than 250 publications in the field of signal processing and pattern recognition. Most of his work deals with automatic speech recognition, where he is particularly concerned with classifier optimization. He also maintains active research programs in vision-based pattern recognition.

Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening

Martin Wöllmer, *Member, IEEE*, Björn Schuller, *Member, IEEE*, Florian Eyben, *Member, IEEE*, and Gerhard Rigoll, *Senior Member, IEEE*

***Abstract*—The automatic estimation of human affect from the speech signal is an important step towards making virtual agents more natural and human-like. In this paper, we present a novel technique for incremental recognition of the user's emotional state as it is applied in a sensitive artificial listener (SAL) system designed for socially competent human–machine communication. Our method is capable of using acoustic, linguistic, as well as long-range contextual information in order to continuously predict the current quadrant in a two-dimensional emotional space spanned by the dimensions valence and activation. The main system components are a hierarchical dynamic Bayesian network (DBN) for detecting linguistic keyword features and long short-term memory (LSTM) recurrent neural networks which model phoneme context and emotional history to predict the affective state of the user. Experimental evaluations on the SAL corpus of non-prototypical real-life emotional speech data consider a number of variants of our recognition framework: continuous emotion estimation from low-level feature frames is evaluated as a new alternative to the common approach of computing statistical functionals of given speech turns. Further performance gains are achieved by discriminatively training LSTM networks and by using bidirectional context information, leading to a quadrant prediction F1-measure of up to 51.3 %, which is only 7.6 % below the average inter-labeler consistency.**

***Index Terms*—Dynamic Bayesian networks (DBNs), emotion recognition, intelligent environments, long short-term memory (LSTM), recurrent neural nets, virtual agents.**

I. Introduction

FOR the design of intelligent environments which enable natural human–machine interaction it is important to consider the principles of interhuman communication as the ideal prototype [1]. While automatic speech recognition (ASR) is already an integral part of most intelligent systems such as virtual agents, in-car interfaces, or mobile phones, a lot more pattern recognition modules are needed to close or at least narrow the gap between the human ability to permanently observe and react to the affective state of the conversational partner in a socially competent way, and the straightforwardness of system responses generated by today's state-of-the-art human–computer interfaces [2], [3]. Therefore, automatic emotion recognition (AER) is an essential precondition to make, e.g., virtual agents more human-like and to increase their acceptance among potential users [4]–[7].

Even though researchers report outstanding recognition accuracies when trying to assign an affective state to an emotionally colored speech turn [8], [9], systems that apply automatic emotion recognition still are only rarely found in every day life. The main reason for this is that emotion recognition performance is often overestimated: apart from examples such as call-center data [10]–[12], databases for interest recognition [13], [14], or other spontaneous speech evaluations [15]–[19], most speech-based AER systems are trained and tested on corpora that contain segmented speech turns with acted, prototypical emotions that are comparatively easy to assign to a set of predefined emotional categories [20]–[22]. Often, only utterances that have been labeled equally by the majority of annotators are used to evaluate AER performance. Yet, these assumptions fail to reflect the conditions a recognition system has to face in real-life usage. Next-generation AER systems must be able to deal with non-prototypical speech data and have to continuously process naturalistic and spontaneous speech as uttered by the user (e.g., as in the Interspeech 2009 Emotion Challenge [23]). More specifically, a real-life emotion recognition engine has to model "everything that comes in," which means it has to use all data as recorded, e.g., for a dialogue system, media retrieval, or surveillance task by using an *open microphone* setting. According to [24], dealing with non-prototypicality is "one of the last barriers prior to integration of emotion recognition from speech into real-life technology."

Thus, in this paper we present and investigate a speech-based system for emotion recognition that is able to cope with spontaneous, non-prototypical, and unsegmented speech. We address the problem of predicting the *quadrant* of an emotional space (spanned by the two dimensions *valence* and *activation*), which best describes the current affective state of the speaker. We willfully omit *dominance* as a further dimension, since we found that activation and dominance are usually strongly correlated. Consequently, the continuum of emotional states is reduced to the four quadrants which can be described as *relaxed/serene* (I), *happy/excited* (II), *sad/bored* (III), and *angry/anxious* (IV) in order to keep the affective state information as simple as

Manuscript received nulldate; revised nulldate; accepted nulldate. Date of publication July 12, 1010; date of current version nulldate. The work was supported by the European Community's Seventh Framework Program (FP7/2007-2013) under Grant 211486 (SEMAINE). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sadaoki Furui.

The authors are with the Institute of Human-Machine-Communication, Technische Universität München, 80333 München, Germany (e-mail: woellmer@tum.de; schuller@tum.de; eyben@tum.de; rigoll@tum.de).

Digital Object Identifier 10.1109/JSTSP.2010.2057200

possible. A further motivation for quadrant quantization of the continuous emotional space is to reduce the multiplicity of possible system responses for the emotion dependent dialogue management of virtual agents, since at some stage, a categorical decision about the user's emotion has to be made before determining a suitable system output. The outlined AER framework is optimized for usage within virtual agent scenarios such as the SEMAINE system for *Sensitive Artificial Listening* [25], which demands for incremental real-time emotion estimation. Applications like the SEMAINE system require customized and immediate feedback based on the emotional state of the user, and responses have to be prepared already before the user has finished speaking. This, however, would hardly be feasible using traditional static classification approaches like support vector machines (SVMs) which classify segmented or fixed length speech segments at the end of a speech turn. Instead, incremental processing demands for techniques that operate on short speech segments while incorporating an adequate and gradually increasing amount of contextual information.

As shown in [26], capturing temporal long-range dependencies is essential for the prediction quality of an AER system and is superior to static SVM modeling. Hence, our technique applies long short-term memory (LSTM) recurrent neural networks [27] which have shown excellent performance in many machine learning applications [28]–[30]. This concept is able to model *emotional history* and overcomes the so-called *vanishing gradient problem* in conventional recurrent neural nets (RNNs). We show that LSTM enables a completely novel approach towards RNN based affect recognition, using low-level features on a frame basis instead of turnwise computed statistical functionals or fixed-length feature vector sequences, as applied in other context-independent RNN systems [31]. Our principle of framewise emotion estimation is related to strategies for speech recognition, where the temporal evolution of low-level descriptors is not only captured by functionals of *features* but by the *classifier*. Such an approach has many advantages: it allows for incremental real-time emotion estimation from speech as it is needed for emotionally sensitive virtual agents and does not need to operate on supra-segmental units of speech (as in almost any other method [32]–[34]). Moreover, the precondition of perfect segmentation is not needed anymore and the AER system can update the emotion prediction *while* the user is speaking. The long short-term memory RNN architecture copes with the fact that speech emotion is a phenomenon observed over a longer time window. Typical units of analysis for static classifiers are complete sentences, sentence fragments (i.e., chunks), or words [35]. Yet, finding the optimal unit of analysis is still an active area of research [9], [36], [37]. Unlike hidden Markov model (HMM)-based methods [38], [39] which also focus on low-level features and perform best-path decoding on the complete input fragment, our technique offers the great advantage that the *amount* of contextual information that is used for emotion recognition is learned during training. In order to refine and update the estimation of a user's emotion once the complete spoken utterance is available, we also investigate the usage of *bidirectional* context [40]. This is done by bidirectional long short-term memory (BLSTM) networks which process the entire speech sequence in forward and backward direction using two hidden layers that are connected to the same output layer. In contrast to the bidirectional system which presumes either offline operation or a short "look-ahead" input buffer, the unidirectional LSTM system can operate in real-time at a moderate computational cost (see Section II.B).

In addition to the acoustic features, the system presented herein also uses linguistic features derived from a dynamic Bayesian network (DBN) for keyword spotting. The DBN is designed in a way that it detects keywords which are correlated to the user's emotion in order to provide a binary linguistic feature vector. In order to also exploit the principle of LSTM modeling for the generation of linguistic features, our system contains an additional LSTM network that provides a discrete phoneme prediction feature to the keyword spotter. This principle of tandem LSTM-DBN modeling was shown to prevail over conventional hidden Markov model-based approaches [41].

The emotion recognition system presented in this paper is trained and evaluated on the Sensitive Artificial Listener (SAL) database [42] which contains natural, spontaneous, and emotionally colored speech. We investigate the accuracy of predicting the quadrants of the emotional space as well as the ability to distinguish high from low activation and valence, respectively. Furthermore, we evaluate the AER performance when considering *neutrality* as a fifth emotional state. We consider both turnwise and framewise classification using BLSTM, LSTM, SVM, and conventional RNN architectures—with and without linguistic features. In addition to continuously estimating valence and activation before assigning the prediction to one of the four quadrants, we also investigate discriminative training on the quadrants.

The rest of this paper is structured as follows. Section II describes the SAL database and gives an overview over the introduced AER system architecture. In Section III, the principle of long short-term memory is introduced. Sections IV and V outline the acoustic and the linguistic feature extractor, respectively. We present experimental results in Section VI and concluding remarks are given in Section VII.

II. Sensitive Artificial Listening

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. This section describes the SAL database which was recorded during a Wizard-of-Oz SAL scenario and will be used in the experimental section of this paper. Further, our AER system architecture will be explained.

A. Database

The SAL corpus is a subset of the HUMAINE database[2] [42] that is continuously labeled in a two-dimensional emotional space spanned by activation and valence. It contains 25 audio-visual recordings in total from four speakers (two male, two female) with an average recording length of 20 minutes per speaker. The language spoken in the database is English.

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

[2]http://emotion-research.net/download/pilot-db/

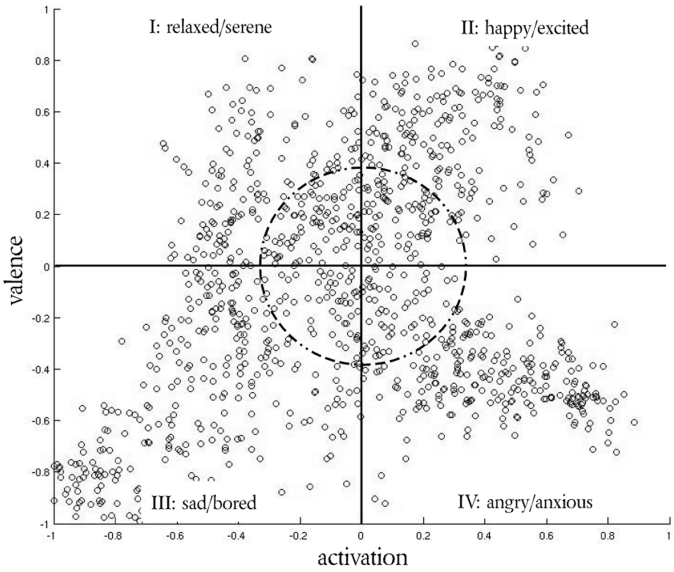

Fig. 1. Turnwise annotations of the SAL database.

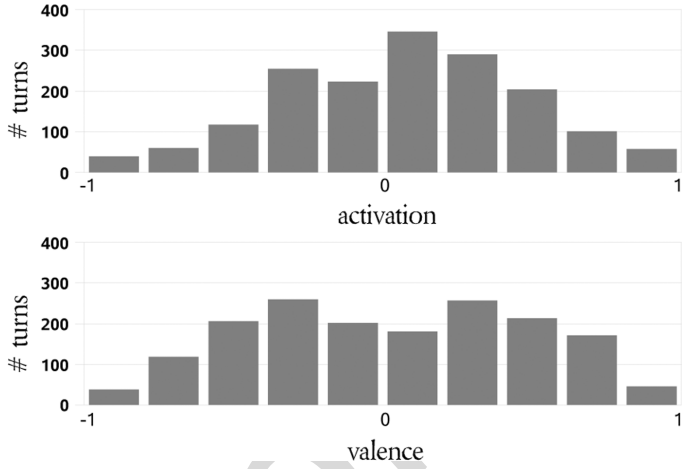

Fig. 2. Histogram for the turnwise annotations of activation (top) and valence (bottom) in the SAL database.

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. All users had to speak to four different virtual characters, each of whom represents one of the four emotional quadrants (Fig. 1): "Prudence" is matter-of-fact (quadrant I), "Poppy" is cheerful (quadrant II), "Obadiah" is pessimistic (quadrant III), and "Spike" is aggressive (quadrant IV). During the conversations, all virtual characters aimed to induce an emotion that corresponds to "their" quadrant. Yet, those "prototypical" virtual characters are used explicitly for emotion induction and not for modeling conditional dependencies between the affective state of the agent and the user, as done in [43] for example. Both, the database and the recording procedure are described in more detail in [42].

The annotators used the FEELtrace system [44] which generates quasi-time-continuous samples of activation and valence every 10 ms (unlike the VAM corpus [45] and practically any other database where labels for the emotional dimensions are given only once per speech turn). All labelers listened to the recordings twice, while annotating activation and valence consecutively in real-time. As ground truth for our experiments, the mean of the four different annotators was used. The mean was calculated by averaging both the (linear) activation and valence coordinates of the labelers for every time step. Note that ambiguous speech turns can lead to the case that the averaged coordinates in the valence-activation space are located in a quadrant that neither of the labelers had assigned to the speech fragment (e.g., the average of coordinates in quadrant I and IV can be located in quadrant II or III). Yet, the resulting quadrant can be seen as the best possible compromise with respect to the average perceived level of activation and valence. An alternative would be to map such ambiguous utterances to a "garbage class." However, since we found that only 2% of the resulting quadrant labels are located in a quadrant that neither of the annotators assigned to the corresponding speech turn, and since *all* of those cases have averaged coordinates that are located in the "neutral" region (coordinates within the dashed circle in Fig. 1), we decided that modeling neutrality is more adequate, rather than the introduction of a "garbage class."

For all experiments reported on in this paper the same training- and test-set splits as introduced in [26] are used. The 25 recording sessions are split into 16 training sessions and nine test sessions. The test split has a total length of 53.3 min, whereas the training split has a length of 99.2 min. Since only four speakers are contained in this database, the training- and test-splits are not speaker disjunctive. Yet, speaker dependent emotion recognition is of significant practical importance, especially for the paradigm of virtual agents and sensitive listeners, since the listener can adapt its models to the current speaker and learn speaker profiles.

For our experiments on turn-based emotion recognition, the sessions were split into turns using an energy based voice activity detection. A total of 1692 turns is accordingly contained in the database. The training- and test splits contain 1102 and 590 turns, respectively. The obtained speech turns do not necessarily comprise complete sentences since the sessions were also split at short hesitation pauses. Thus, the average length of a speech turn is 3.5 seconds. Since the turns are short enough to assume quasi-stationarity of the emotion within a turn, labels for each turn were computed by averaging the FEELtrace annotations for valence and activation over a complete turn in order to obtain a ground truth for the turnwise AER experiments. Note that, unlike in databases annotated on the word level [15], short "activation peaks" like the stress of a single word within a sentence are unlikely to be captured by the annotators, due to the finite reaction time of the human labelers. Consequently, the time-continuous annotations tend to have low-pass characteristics and do not contain high frequencies, which limits the loss of information due to the averaging of annotation samples within a turn and accounts for the fact that emotion is perceived over a longer time window. The distribution of the averaged labels can be seen in Figs. 1 and 2. The dashed circle (with a radius of 0.33, dividing the axes into thirds) in the center of the valence-activation space in Fig. 1 marks a fifth region which represents a neutral emotional state. The coordinates that lie within this circle will be considered as belonging to a fifth, neutral class (see Section VI).

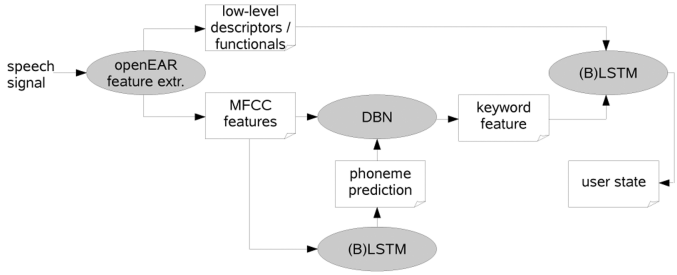

Fig. 3. Architecture of the acoustic-linguistic affect recognition system.

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 preselected *emotional prototypes* as in virtually any other database. Note that there is usually a high difference in accuracy between the tasks of prototypical and non-prototypical emotion recognition [23], [24], [46].

B. System Architecture

In Fig. 3, a flowchart of the presented incremental affect recognition system is shown. Processing components such as the LSTM network or the feature extractors are represented as ovals, whereas rectangles denote data. Depending on whether framewise or turnwise processing is used, our openEAR feature extractor module [47] (see Section IV) provides either low-level descriptors or statistical functionals of acoustic low-level features to the LSTM network (outlined in Section III) for emotion estimation. Additionally, mel-frequency cepstral coefficient (MFCC) features are provided to both components of the tandem keyword spotter component (see Section V), consisting of a DBN and a further LSTM network for phoneme prediction. Together with the produced phoneme predictions, the MFCC features are observed by the DBN, which then can detect the occurrence of a relevant keyword (i.e., a word that is relevant for valence or activation prediction, see Section V). Both, the discrete keyword feature and the acoustic features extracted by openEAR are used by an LSTM network to predict the user's current emotion. For the emotion coding, EmotionML[3] is used [48], [49], supporting continuous spatio-temporal emotion representation. EmotionML is a standard representation format for emotion-related states in technological contexts, developed by the W3C Emotion Markup Language Incubator Groups. It can be used within the tasks of data annotation, emotion recognition, and generation of emotion-related states.

Details about the overall architecture of the SEMAINE dialogue system can be found in [25].

Due to the complexity of the system, the computational cost of our AER engine is higher than for standard classification techniques such as SVMs, which however show significantly lower performance than the proposed system (see Section VI). Yet, when exclusively using *unidirectional* context within the LSTM framework, the causal system can operate in real-time: on an AMD Phenom 64 bit quad core CPU at 2.2 GHz, the openEAR feature extraction module runs online with a real-time factor (RTF) of 0.01, while the LSTM operates at a real-time factor of 0.09. Only one of the four cores was used for computation. Time and space complexity of the DBN is $\mathcal{O}(T \log T)$ and $\mathcal{O}(\log T)$, respectively, assuming that T corresponds to the length of the speech sequence that is currently processed.

[3]http://www.w3.org/2005/Incubator/emotion/XGR-emotionml-20081120/

III. Long Short-Term Memory

This section outlines the principle of the long short-term memory RNNs that are used for emotion classification in Section VI as well as for phoneme prediction in Section V. Framewise classification of emotion as investigated in this paper presumes a classifier that can access and model long-range context, since emotion mostly affects the long-term *dynamics* of prosodic, spectral, and voice quality speech features. When attempting to predict emotion frame by frame, a large number of preceding speech frames have to be taken into account in order to capture speech characteristics that are influenced by emotion. The *number* of speech frames which should be used to obtain enough context for reliably estimating emotion without affecting the capability of also detecting sudden changes of the speaker's emotional state is hard to determine [36], [37]. Thus, a classifier that is able to *learn* the amount of context is a promising alternative to manually defining fixed time windows for emotion recognition. Static techniques such as SVMs do not explicitly model context but rely on either capturing contextual information via statistical functionals of features [14] or aggregating frames using multi-instance learning techniques [50]. Dynamic classifiers like hidden Markov models are often used for flexible context modeling and time warping. Yet, HMMs have drawbacks such as the inherent assumption of conditional independence of successive observations, meaning that an observation is statistically independent of past observations provided that the values of the hidden variables are known. Hidden conditional random fields (HCRFs) [51] are one attempt to overcome this limitation. However, HCRF also offer no possibility to model a self-learned amount of contextual information. Other dynamic classifiers such as neural networks are able to model a certain amount of context by using cyclic connections. These so-called recurrent neural networks can in principle map from the entire *history* of previous inputs to each output. Yet, the analysis of the error flow in conventional recurrent neural nets led to the finding that long range context is inaccessible to standard RNNs since the backpropagated error either blows up or decays over time (vanishing gradient problem [52]). This led to the introduction of long short-term memory RNNs [27]. They are able to overcome the vanishing gradient problem and can learn the optimal amount of contextual information relevant for the classification task. Thus, LSTM architectures seem to be well-suited for our framewise emotion recognition task.

An LSTM layer is composed of recurrently connected memory blocks, each of which contains one or more memory cells, along with three multiplicative "gate" units: the input, output, and forget gates. The gates perform functions analogous to read, write, and reset operations. More specifically, the cell input is multiplied by the activation of the input gate, the cell output by that of the output gate, and the previous cell values by the forget gate (see Fig. 4). The overall effect is to allow the network to store and retrieve information over long periods of

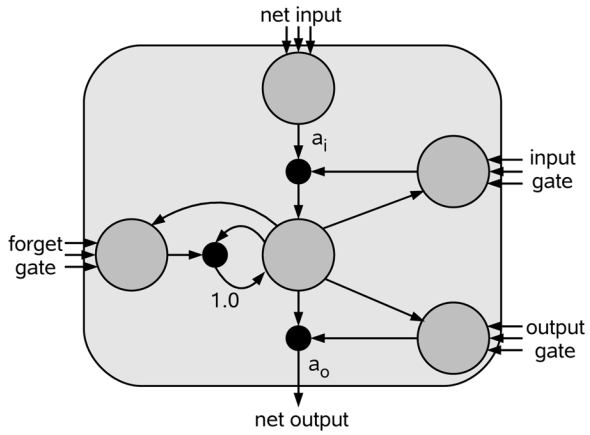

Fig. 4. LSTM memory block consisting of one memory cell: the input, output, and forget gates collect activations from inside and outside the block which control the cell through multiplicative units (depicted as small circles); input, output, and forget gate scale input, output, and internal state, respectively; a_i and a_o denote activation functions; the recurrent connection of fixed weight 1.0 maintains the internal state.

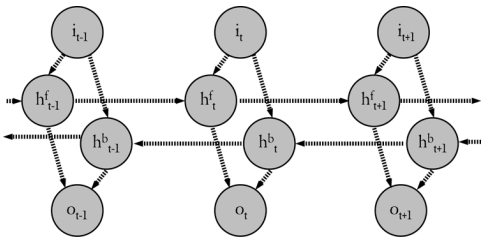

Fig. 5. Structure of a bidirectional network with input i, output o, and two hidden layers (h^f and h^b) for forward and backward processing.

TABLE I
28 LOW-LEVEL AUDIO FEATURES FOR TIME-CONTINUOUS EMOTION ANALYSIS (C) AND 39 FEATURES FOR TURN-BASED RECOGNITION (T); FEATURES IN BOLD FACE ARE USED FOR BOTH, CONTINUOUS AND TURN-BASED RECOGNITION

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 0, **MFCC 1-12**, MFCC 13-15	12	16
Time Signal	**Zero-Crossing-Rate**, max. and min. value, DC component	1	4
Spectral	**Energy in bands 0-250Hz, 0-650Hz, 250-650Hz, 1000-4000Hz**	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**

time. For example, as long as the input gate remains closed, the activation of the cell will not be overwritten by new inputs and can therefore be made available to the net much later in the sequence by opening the output gate.

Another problem with standard RNNs is that they have access to past but not to future context. This can be overcome by using bidirectional RNNs [40], where two separate recurrent hidden layers scan the input sequences in opposite directions. The two hidden layers are connected to the same output layer, which therefore has access to context information in both directions. The amount of context information that the network actually uses is learned during training, and does not have to be specified beforehand. Fig. 5 shows the structure of a simple bidirectional network.

Combining bidirectional networks with LSTM gives bidirectional LSTM [53], which has demonstrated excellent performance in phoneme recognition [28], [54], keyword spotting [29], and emotion recognition from speech [26].

While bidirectional LSTM cannot be used for online incremental prediction tasks, they are well suited to refine or correct the estimation of affect once the complete turn is available. Thus, we included bidirectional networks in our performance evaluation on the SAL database.

All RNN-based classifiers used in the experiments in Section VI were implemented using the open source RNNLIB library.[4]

TABLE II
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

IV. ACOUSTIC FEATURE EXTRACTION

Acoustic features from the speech signal are extracted using our openEAR [47] audio feature extractor, which was also used to provide features for the Interspeech 2009 Emotion Challenge [23].

The 28 low-level descriptors extracted from the audio signal for time-continuous emotion recognition are summarized in Table I (column "C'). The descriptors were extracted every 20 ms for overlapping frames with a frame-length of 32 ms. First-order regression coefficients are appended to the 28 low-level descriptors, resulting in a 56-dimensional feature vector for each frame.

In order to enable also turn-based emotion recognition experiments, the openEAR module alternatively follows the traditional approach of generating a large set of features by applying statistical functionals to low-level descriptor contours. An extended set of 39 low-level-descriptors detailed in Table I (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, resulting in a total of 4212 features. The 36 functionals are detailed in Table II.

The 4212 features for turn-based emotion recognition are reduced to relevant features for activation and valence independently by a correlation-based feature subset (CFS) selection

[4]http://github.com/alexgraves/RNNLIB

[55], [56]. The main idea of CFS is that useful feature subsets should contain features that are highly correlated with the target class while being uncorrelated with each other. The core of CFS is an evaluation function

$$M_S = \frac{k \cdot r_{\mathrm{cf}}}{\sqrt{k + k(k-1) r_{\mathrm{ff}}}} \quad (1)$$

where M_S is the rating of a subset S with k features. r_{cf} denotes the mean feature-class correlation and r_{ff} is the average feature-feature inter-correlation. Good subsets of features have highly predictive properties, yielding a high value in the numerator of (1), and a low degree of redundancy among the features, yielding a small value in the denominator. For correlation measurement, the symmetrical uncertainty coefficient is used (as described in [55]). To avoid an exhaustive search in the feature space a greedy hill climbing forward search is applied [56]. In this heuristic search algorithm, each feature is tentatively added to the feature subset, whereas the resulting set of features is evaluated using (1). Once the (so far) best feature set has been chosen, the procedure is repeated. Note that we willfully decided for a filter based feature selection method, since a wrapper-based technique would have biased the resulting feature set with respect to compatibility to a specific classifier.

Conducting CFS for turn-based emotion recognition via regression resulted in 60 features being selected for activation and 64 features for valence[5]. As termination criterion we considered a maximum of five non-improving nodes before terminating the greedy hill climbing forward search. Binary targets for activation and valence (high versus low, see Section VI) lead to the selection of 110 and 55 features, respectively. For the discriminative four-class quadrant classification task 121 features were selected, and for the five-class task applying CFS resulted in 123 selected features. Framewise emotion recognition uses the full set of $28 \cdot 2 = 56$ features without further reduction.

All features (turn-based functionals and low-level features) were standardized to have zero mean and unit standard deviation. These parameters were computed from the training data only and applied to both training and test data.

V. Linguistic Feature Extraction

This section outlines the tandem LSTM-DBN keyword spotter which generates binary linguistic features in order to incorporate knowledge about the spoken content via early fusion.

A. Background and References

Apart from acoustic features, also spoken or written text carries information about the underlying affective state [57]–[59]. This is usually reflected in the usage of certain words or grammatical alterations. A number of approaches exist for this analysis: keyword spotting [60], [61], rule-based modeling [62], semantic trees [63], Latent Semantic Analysis [64], transformation-based learning [65], world-knowledge-modeling [66], key-phrase spotting [67], and Bayesian networks [68], [69]. Two methods seem to be predominant, presumably because they are shallow representations of linguistic knowledge and have already been frequently employed in automatic speech processing: (class-based) N-grams [70]–[73] and vector space modeling [74], [75]. Due to the typical data sparseness in emotion recognition, unigrams mostly have been applied so far [72], [73]. The technique applied in our experiments is related to bag of words modeling [74]–[76] via keyword spotting; however, when applying framewise emotion recognition, only one keyword can be present at a given time frame. In the case of turnwise AER, the binary feature vector can contain more than one keyword. This would enable techniques like (bag of) N-gram modeling or other forms of linguistic information integration [77], [78], which however were not conducted in this paper in order to allow a fair comparison between framewise and turnwise affect recognition.

For combined acoustic and linguistic AER, the acoustic feature vector is extended by appending binary linguistic features. Each binary feature corresponds to the occurrence of one of the 56 keywords that were shown to be correlated to either valence or activation. 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."

When applying framewise acoustic-linguistic analysis, a short buffer 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 as linguistic features can still be delivered while the user is speaking. In order to reduce the vocabulary to a small set of emotionally meaningful keywords, correlation-based feature subset selection was applied on the training set. Pace regression [79]-based CFS used the continuous labels for valence and activation for bag of words keyword selection with a minimum term frequency of two (without stemming). Thereby 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*.[5]

The keyword spotter used in this paper is based on a recently introduced hierarchical DBN which was shown to significantly outperform a standard HMM-based approach [80]. The incorporation of an LSTM layer providing improved phoneme predictions was proven to further enhance keyword detection performance [41].

B. Design Overview

The tandem LSTM-DBN architecture we used for keyword spotting was proven to be robust with respect to phoneme recognition errors [41] and well suited for emotional speech. Its structure is depicted in Fig. 6. The network is composed of five different layers and hierarchy levels, respectively: a word layer, a phoneme layer, a state layer, the observed features, and the LSTM layer, consisting of inputs i_t, a hidden layer h_t, and outputs o_t (nodes inside the grey shaded box).

The following random variables are defined for every time step t: q_t denotes the phoneme identity, q_t^{ps} represents the position within the phoneme, q_t^{tr} indicates a phoneme transition,

[5]an explanation of the used features, openEAR configuration files, and lists of the selected features and keywords can be found at http://www.openaudio.eu/features_emo09.zip

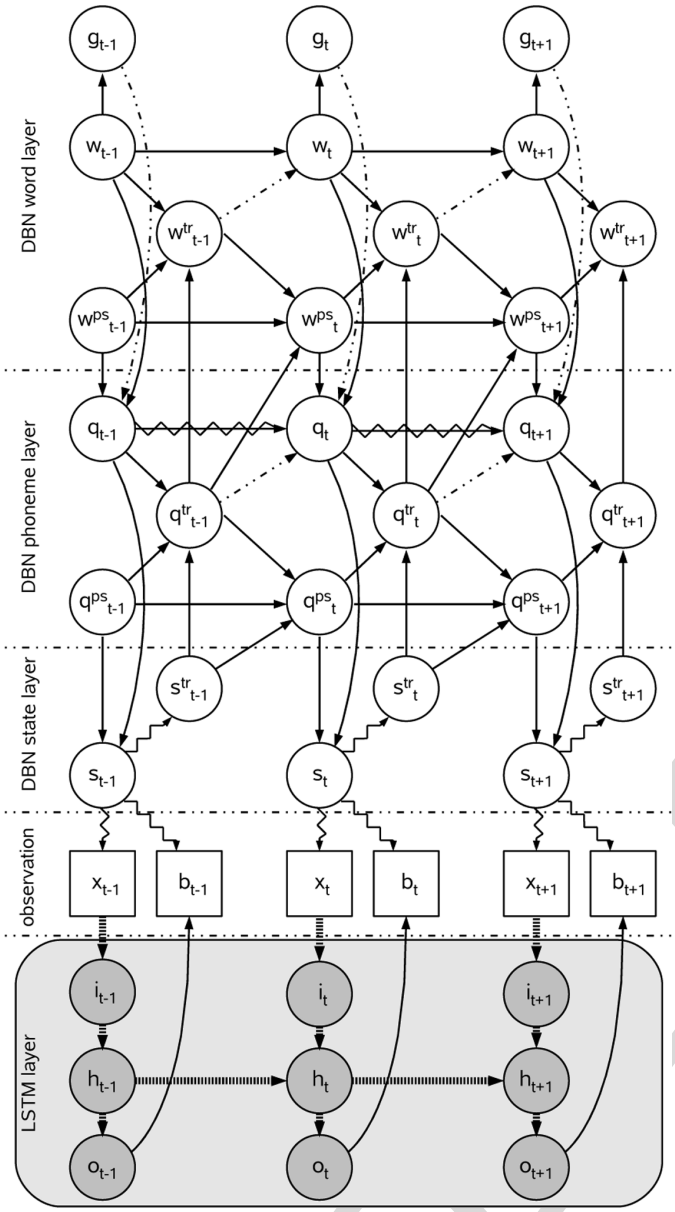

Fig. 6. Structure of the tandem LSTM-DBN keyword spotter: the LSTM network (gray shaded box) provides a discrete phoneme prediction feature b_t which is observed by the DBN, in addition to the MFCC features x_t. The DBN is composed of a state, phoneme, and word layer, consisting of hidden transition $(s_t^{\mathrm{tr}}, q_t^{\mathrm{tr}}, w_t^{\mathrm{tr}})$, position $(q_t^{\mathrm{ps}}, w_t^{\mathrm{ps}})$, and identity (s_t, q_t, w_t) variables. Hidden variables (circles) and observed variables (squares) are connected via random CPFs (zig-zagged lines) or deterministic CPFs (straight lines). Switching parent dependencies are indicated with dotted lines.

s_t is the current state with s_t^{tr} indicating a state transition, and x_t denotes the observed MFCC features. The variables w_t, w_t^{ps}, and w_t^{tr} are identity, position, and transition variables for the word layer of the DBN whereas a hidden *garbage variable* g_t indicates whether the current word is a keyword or not. A second observed variable b_t contains the phoneme prediction of the LSTM network. Fig. 6 displays hidden variables as circles and observed variables as squares. Deterministic conditional probability functions (CPFs) are represented by straight lines and zig-zagged lines correspond to random CPFs. Dotted lines refer to so-called *switching parents* [81], which allow a variable's parents to change conditioned on the current value of the switching parent. Note that a switching parent can not only change the set of parents but also the implementation (i.e., the CPF) of a parent. The bold dashed lines in the LSTM layer do not represent statistical relations but simple data streams.

C. Design Details

Assuming a speech sequence of length T, the DBN structure specifies the factorization

$$
\begin{aligned}
&p\left(g_{1:T}, w_{1:T}, w_{1:T}^{\mathrm{tr}}, w_{1:T}^{\mathrm{ps}}, q_{1:T}, q_{1:T}^{\mathrm{tr}}, q_{1:T}^{\mathrm{ps}}, s_{1:T}^{\mathrm{tr}}, s_{1:T}, x_{1:T}, b_{1:T}\right) \\
&\quad = \prod_{t=1}^{T} p(x_t \mid s_t) p(b_t \mid s_t) f\left(s_t \mid q_t^{\mathrm{ps}}, q_t\right) \\
&\qquad \times p\left(s_t^{\mathrm{tr}} \mid s_t\right) f\left(q_t^{\mathrm{tr}} \mid q_t^{\mathrm{ps}}, q_t, s_t^{\mathrm{tr}}\right) f(g_t \mid w_t) \\
&\qquad \times f\left(w_t^{\mathrm{tr}} \mid q_t^{\mathrm{tr}}, w_t^{\mathrm{ps}}, w_t\right) f\left(q_1^{\mathrm{ps}}\right) p\left(q_1 \mid w_1^{\mathrm{ps}}, w_1, g_1\right) \\
&\qquad \times f\left(w_1^{\mathrm{ps}}\right) p(w_1) \prod_{t=2}^{T} f\left(q_t^{\mathrm{ps}} \mid s_{t-1}^{\mathrm{tr}}, q_{t-1}^{\mathrm{ps}}, q_{t-1}^{\mathrm{tr}}\right) \\
&\qquad \times p\left(w_t \mid w_{t-1}^{\mathrm{tr}}, w_{t-1}\right) \\
&\qquad \times p\left(q_t \mid q_{t-1}^{\mathrm{tr}}, q_{t-1}, w_t^{\mathrm{ps}}, w_t, g_t\right) \\
&\qquad \times f\left(w_t^{\mathrm{ps}} \mid q_{t-1}^{\mathrm{tr}}, w_{t-1}^{\mathrm{ps}}, w_{t-1}^{\mathrm{tr}}\right)
\end{aligned} \tag{2}
$$

with $p(\cdot)$ denoting random conditional probability functions and $f(\cdot)$ describing deterministic CPFs.

The probability of the observed sequence can then be computed by summing over all hidden variables, whereas the factorization property in (2) can be exploited to optimally distribute the sums over the hidden variables into the products, using the junction tree algorithm [82].

The size of the LSTM input layer i_t corresponds to the dimensionality of the acoustic feature vector x_t, whereas the vector o_t contains one probability score for each of the P different phonemes at each time step. b_t is the index of the most likely phoneme:

$$
b_t = \max_{o_t}(o_{t,1}, \ldots, o_{t,j}, \ldots, o_{t,P}). \tag{3}
$$

The CPFs $p(x_t \mid s_t)$ are described by Gaussian mixtures, as is common practice with HMMs. Together with $p(b_t \mid s_t)$ and $p(s_t^{\mathrm{tr}} \mid s_t)$, they are learned via EM training. s_t^{tr} is a binary variable, indicating whether a state transition takes place or not. Since the current state is known with certainty, given the phoneme and the phoneme position, $f(s_t \mid q_t^{\mathrm{ps}}, q_t)$ is purely deterministic. A phoneme transition occurs whenever $s_t^{\mathrm{tr}} = 1$ and $q_t^{\mathrm{ps}} = S$ provided that S denotes the number of states of a phoneme. This is expressed by the function $f(q_t^{\mathrm{tr}} \mid q_t^{\mathrm{ps}}, q_t, s_t^{\mathrm{tr}})$. The phoneme position q_t^{ps} is known with certainty if s_{t-1}^{tr}, q_{t-1}^{ps}, and q_{t-1}^{tr} are given.

The hidden variable w_t can take values in the range $w_t = 0 \ldots K$ with K being the number of different keywords in the vocabulary. In case $w_t = 0$, the model is in the *garbage state* which means that no keyword is uttered at that time. The variable g_t is then equal to one. w_{t-1}^{tr} is a switching parent of w_t: if no word transition is indicated, w_t is equal to w_{t-1}. Otherwise, a word bigram specifies the CPF $p(w_t \mid w_{t-1}^{\mathrm{tr}} = 1, w_{t-1})$. In our experiments, we simplified the word bigram to a zerogram

which makes each keyword equally likely. However, we introduced differing *a priori* likelihoods for keywords and garbage phonemes:

$$p\left(w_t = 1:K \mid w_{t-1}^{\mathrm{tr}} = 1\right) = \frac{K \cdot 10^a}{K \cdot 10^a + 1} \tag{4}$$

and

$$p\left(w_t = 0 \mid w_{t-1}^{\mathrm{tr}} = 1\right) = \frac{1}{K \cdot 10^a + 1}. \tag{5}$$

The parameter a can be used to adjust the tradeoff between true positives and false positives. Setting $a = 0$ means that the *a priori* probability of a keyword and the probability that the current phoneme does not belong to a keyword are equal. Adjusting $a > 0$ implies a more aggressive search for keywords, leading to higher true positive and false positive rates. The CPFs $f(w_t^{\mathrm{tr}} \mid q_t^{\mathrm{tr}}, w_t^{\mathrm{ps}}, w_t)$ and $f(w_t^{\mathrm{ps}} \mid q_{t-1}^{\mathrm{tr}}, w_{t-1}^{\mathrm{ps}}, w_{t-1}^{\mathrm{tr}})$ are similar to the phoneme layer of the DBN (i.e., the CPFs for q_t^{tr} and q_t^{ps}). However, we assume that "garbage words" always consist of only one phoneme, meaning that if $g_t = 1$, a word transition occurs as soon as $q_t^{\mathrm{tr}} = 1$. Consequently, w_t^{ps} is always zero if the model is in the garbage state. The variable q_t has two switching parents: q_{t-1}^{tr} and g_t. Similar to the word layer, q_t is equal to q_{t-1} if $q_{t-1}^{\mathrm{tr}} = 0$. Otherwise, the switching parent g_t determines the parents of q_t. In case $g_t = 0$—meaning that the current word is a keyword—q_t is a deterministic function of the current keyword w_t and the position within the keyword w_t^{ps}. If the model is in the garbage state, q_t only depends on q_{t-1} in a way that phoneme transitions between identical phonemes are forbidden.

Note that the design of the CPF $p(q_t \mid q_{t-1}^{\mathrm{tr}}, q_{t-1}, w_t^{\mathrm{ps}}, w_t, g_t)$ entails that the DBN will strongly tend to choose $g_t = 0$ (i.e., it will detect a keyword) once a phoneme sequence that corresponds to a keyword is observed. Decoding such an observation while being in the garbage state $g_t = 1$ would lead to "phoneme transition penalties' since the CPF $p(q_t \mid q_{t-1}^{\mathrm{tr}} = 1, q_{t-1}, w_t^{\mathrm{ps}}, w_t, g_t = 1)$ contains probabilities less than one. By contrast, $p(q_t \mid q_{t-1}^{\mathrm{tr}} = 1, w_t^{\mathrm{ps}}, w_t, g_t = 0)$ is deterministic, introducing no likelihood penalties at phoneme borders.

The DBN was implemented using the Graphical Models Toolkit (GMTK) [83]. In our experiments, we used phoneme models consisting of three states with 16 Gaussian mixtures. Phoneme models were trained on the TIMIT database [84] and adapted using the training split of the Sensitive Artificial Listener database (see Section II-A) to allow a better modeling of emotionally colored speech. Thereby all means, variances, and weights of the Gaussian mixture probability distributions $p(x_t \mid s_t)$, as well as the state transition probabilities $p(s_t^{\mathrm{tr}} \mid s_t)$ were re-estimated until the change of the overall log likelihood of the SAL training set became less than 0.02%. Since we found that in the context of our target application a low true positive rate is less critical than a high false positive rate, we chose a low tradeoff parameter of $a = 0$. The LSTM network of the tandem keyword spotter consists of 100 memory blocks of one cell each. All other DBN and LSTM parameters correspond exactly to those applied in [41]. Using these settings, the keyword spotter achieves a true positive rate of 0.59 at a false positive rate of 0.05 on the test partition of the SAL corpus.

VI. Experiments

Our emotion recognition engine was trained and tested on the SAL database (see Section II-A). In order to fit the requirements of the SEMAINE dialogue management [25], the recognition framework was designed in a way that it estimates the current quadrant in the two-dimensional valence-activation space. In addition to quadrant classification, we also investigated a five-class task including a "neutral" state, as well as discriminating low and high valence and activation separately.

A. Primary Systems Evaluated

For quadrant prediction we followed two different strategies: first, we trained LSTM networks for regression to obtain continuous predictions for valence and activation which were then mapped onto one of the four quadrants. In order to conduct feature selection independently for both the valence and the activation dimension, we used separate networks for the two dimensions. Second, the continuous labels for the emotional dimensions were mapped *before* training the network in order to allow a discriminative training on the quadrants, following the strategy introduced in [85]. These two strategies were also evaluated for the five-class task and for both of the two-class tasks (discrimination of low versus high activation and valence, respectively).

For each of the two techniques we evaluated both traditional turnwise classification with statistical functionals of acoustic features (see Section IV) and framewise classification using only low-level features. The gain of appending the binary keyword feature vector obtained by the dynamic Bayesian network (outlined in Section V) for combined acoustic-linguistic affect recognition was examined for every recognizer configuration.

The size of the LSTM input layer corresponds to the number of selected acoustic and linguistic features (see Sections IV and V), while the size of the output layer is equal to the number of regression/classification targets (one, two, four, and five, respectively). Each LSTM-RNN consists of one recurrent hidden layer with 50 memory blocks of one LSTM cell each. The BLSTM-RNN has two hidden layers of 50 memory blocks, one for each direction (forwards, backwards). For the acoustic–linguistic experiments the LSTM network size was increased to 70 memory blocks due to the increased size of the combined acoustic–linguistic feature vector. The networks were trained applying resilient propagation [86]. Prior to training, all weights were randomly initialized in the range from -0.1 to 0.1. Input and output gates used $\tanh$ activation functions, while the forget gates had logistic activation functions. Since the training converged faster for turnwise classification, we aborted turnwise training after ten epochs, whereas the training procedure for framewise classification was aborted after 250 epochs.

Before mapping the (B)LSTM-RNN predictions o_t onto quadrants, they were smoothed using a first-order low-pass filter to obtain the filtered predictions o_t^s

$$o_t^s = \alpha o_{t-1}^s + (1 - \alpha) \cdot o_t. \tag{6}$$

TABLE III
KAPPA VALUES FOR THE FOUR DIFFERENT ANNOTATORS IN THE SAL DATABASE (TURNWISE QUADRANT LABELING); ILA: INTER-LABELER AGREEMENT

κ	1	2	3	4
ILA	0.68	0.67	0.67	0.60
1		0.49	0.48	0.46
2			0.48	0.45
3				0.52

An α of 0.99 was used for time-continuous emotion recognition and an α of 0.7 was used for turn-based recognition. Both values were optimized on the training set.

B. Comparison Systems and Ground Truth

As a common continuous recognition technique, support vector regression (SVR) was performed for comparison [26], [56], [87]. The SVR used a polynomial kernel function of degree 1 and sequential minimal optimization (SMO). The discriminatively trained LSTM networks were compared to SVMs instead of SVR. Since SVR and SVM do not model contextual information, only turnwise classification was evaluated in this case. In order to determine the gain of long short-term memory modeling, we also investigated conventional RNN classification for comparison. The RNNs were trained in the same way as the LSTM networks; however, the network consisted of 50 hidden neurons instead of the 50 one-cell LSTM memory blocks.

Furthermore, we evaluated inter-labeler consistency as an upper benchmark for automatic emotion recognition. To obtain an impression of human emotion prediction quality we compared the annotations of one labeler to the mean of the annotations of the remaining three labelers. This was done for all of the four labelers so that eventually the average inter-labeler consistency could be determined.

As a further evaluation of inter-labeler agreement, Table III shows the kappa values for the four different annotators. Since each of the kappa values is larger than 0.4, the labeler agreement can be characterized as sufficiently high.

C. Results

Tables IV and VI show the recognition result for the assignment of quadrants using the regression method and the discriminative technique, respectively. Results for the five-class task which also considers a "neutral" state (see Fig. 1) can be seen in Tables V and VII, and Tables VIII and IX contain the results for separate classification of the degree of activation and valence (i.e., positive versus negative activation and valence, respectively). Due to the slightly unbalanced class distribution, accuracy is a rather inappropriate performance measure. Thus, we used the F1-measure as the harmonic mean between unweighted recall and unweighted precision for performance evaluation. Compared to emotion recognition on prototypical speech turns (as in [8] or [9]), the overall performance is significantly lower. Yet, the accuracies are in the order of magnitude that is typical for real-life experiments, attempting to classify natural, non-prototypical, and ambiguous emotional speech turns [23].

TABLE IV
REGRESSION-(B)LSTM AND RNN PERFORMANCE, SUPPORT VECTOR REGRESSION (SVR) PERFORMANCE, AND AVERAGE LABELER (LAB) CONSISTENCY FOR QUADRANT CLASSIFICATION USING TURNWISE OR FRAMEWISE PREDICTION WITH ACOUSTIC (A) OR ACOUSTIC–LINGUISTIC (A + L) FEATURES: ACCURACY (ACC.), UNWEIGHTED RECALL (REC.), UNWEIGHTED PRECISION (PREC.), AND F1-MEASURE (F1)

model	unit	features	acc.	rec.	prec.	F1
			quadrants			
BLSTM	turn	A	37.1 %	34.9 %	35.5 %	35.2 %
BLSTM	turn	A+L	41.0 %	36.9 %	37.8 %	37.3 %
BLSTM	frame	A	41.7 %	44.8 %	42.0 %	43.3 %
BLSTM	frame	A+L	48.2 %	51.6 %	49.3 %	**50.4 %**
LSTM	turn	A	37.3 %	37.9 %	35.4 %	36.6 %
LSTM	turn	A+L	38.6 %	38.4 %	39.8 %	39.7 %
LSTM	frame	A	31.2 %	33.4 %	37.2 %	35.2 %
LSTM	frame	A+L	34.2 %	30.7 %	37.9 %	33.9 %
RNN	turn	A	33.7 %	34.8 %	34.7 %	34.7 %
RNN	turn	A+L	37.1 %	35.5 %	36.7 %	36.1 %
RNN	frame	A	31.0 %	36.9 %	33.8 %	35.3 %
RNN	frame	A+L	28.2 %	31.7 %	34.8 %	33.2 %
SVR	turn	A	28.8 %	30.0 %	27.3 %	28.6 %
SVR	turn	A+L	33.3 %	32.2 %	30.4 %	31.3 %
lab	*turn*		*62.0 %*	*59.2 %*	*58.7 %*	*58.9 %*
lab	*frame*		*59.2 %*	*58.3 %*	*56.7 %*	*57.4 %*

TABLE V
REGRESSION-(B)LSTM AND RNN PERFORMANCE, SUPPORT VECTOR REGRESSION (SVR) PERFORMANCE, AND AVERAGE LABELER (LAB) CONSISTENCY FOR QUADRANT/NEUTRAL FIVE-CLASS TASK USING TURNWISE OR FRAMEWISE PREDICTION WITH ACOUSTIC (A) OR ACOUSTIC–LINGUISTIC (A + L) FEATURES: ACCURACY (ACC.), UNWEIGHTED RECALL (REC.), UNWEIGHTED PRECISION (PREC.), AND F1-MEASURE (F1)

model	unit	features	acc.	rec.	prec.	F1
			quadrants + neutral			
BLSTM	turn	A	37.9 %	34.1 %	38.6 %	36.2 %
BLSTM	turn	A+L	40.9 %	30.6 %	39.5 %	34.5 %
BLSTM	frame	A	34.6 %	39.3 %	34.3 %	36.6 %
BLSTM	frame	A+L	44.2 %	49.4 %	45.2 %	**47.2 %**
LSTM	turn	A	36.0 %	35.1 %	32.5 %	33.7 %
LSTM	turn	A+L	39.0 %	30.0 %	35.5 %	32.5 %
LSTM	frame	A	29.0 %	28.3 %	32.5 %	30.3 %
LSTM	frame	A+L	33.2 %	30.4 %	30.3 %	30.4 %
RNN	turn	A	35.1 %	30.9 %	33.2 %	32.0 %
RNN	turn	A+L	36.8 %	30.8 %	34.4 %	32.5 %
RNN	frame	A	35.6 %	21.1 %	41.4 %	27.9 %
RNN	frame	A+L	36.8 %	20.5 %	41.0 %	27.4 %
SVR	turn	A	32.8 %	25.5 %	24.9 %	25.2 %
SVR	turn	A+L	32.0 %	25.2 %	24.9 %	25.0 %
lab	*turn*		*56.8 %*	*55.1 %*	*53.7 %*	*54.3 %*
lab	*frame*		*56.3 %*	*56.9 %*	*54.9 %*	*55.8 %*

A rating of the prediction quality can be obtained when comparing the best result in Table IV (framewise BLSTM classification using acoustic and linguistic features) with the prediction performance of a human labeler (lab, frame in Table IV): when comparing the annotation of a single labeler to the mean of the annotations of the remaining three labelers, the obtained average F1-measure (57.4%) is only 7% higher than the F1-measure of the best classifier (50.4%). This reflects the ambiguity of perceived emotion and the resulting low degree of inter-labeler agreement. A further reason for the low annotator F1-measure is that a high amount of utterances are near the class borders (see Fig. 1). Consequently, those speech turns are hard to assign, even for human annotators. Such non-prototypical, ambiguous utterances also reduce the uncertainty during model training, which limits the obtainable automatic recognition performance.

TABLE VI
DISCRIMINATIVE (B)LSTM AND RNN PERFORMANCE, SUPPORT VECTOR MACHINE (SVM) PERFORMANCE, AND AVERAGE LABELER (LAB) CONSISTENCY FOR QUADRANT CLASSIFICATION USING TURNWISE OR FRAMEWISE PREDICTION WITH ACOUSTIC (A) OR ACOUSTIC–LINGUISTIC (A + L) FEATURES: ACCURACY (ACC.), UNWEIGHTED RECALL (REC.), UNWEIGHTED PRECISION (PREC.), AND F1-MEASURE (F1)

model	unit	features	acc.	rec.	prec.	F1
			quadrants			
BLSTM	turn	A	49.3 %	51.3 %	51.2 %	**51.3 %**
BLSTM	turn	A+L	47.6 %	48.6 %	46.8 %	47.7 %
BLSTM	frame	A	42.5 %	43.9 %	41.3 %	42.5 %
BLSTM	frame	A+L	39.0 %	37.4 %	37.1 %	37.2 %
LSTM	turn	A	48.6 %	47.4 %	48.2 %	47.8 %
LSTM	turn	A+L	44.9 %	49.1 %	48.3 %	48.7 %
LSTM	frame	A	37.4 %	38.0 %	38.1 %	38.1 %
LSTM	frame	A+L	32.0 %	37.8 %	32.6 %	35.3 %
RNN	turn	A	46.3 %	47.2 %	47.2 %	47.2 %
RNN	turn	A+L	45.9 %	46.5 %	45.8 %	46.1 %
RNN	frame	A	28.3 %	32.1 %	30.9 %	31.5 %
RNN	frame	A+L	22.1 %	28.2 %	27.3 %	27.7 %
SVM	turn	A	39.0 %	39.6 %	41.2 %	40.4 %
SVM	turn	A+L	37.8 %	38.5 %	36.7 %	37.6 %
lab	*turn*		*62.0 %*	*59.2 %*	*58.7 %*	*58.9 %*
lab	*frame*		*59.2 %*	*58.3 %*	*56.7 %*	*57.4 %*

TABLE VII
DISCRIMINATIVE (B)LSTM AND RNN PERFORMANCE, SUPPORT VECTOR REGRESSION (SVR) PERFORMANCE, AND AVERAGE LABELER (LAB) CONSISTENCY FOR QUADRANT/NEUTRAL FIVE-CLASS TASK USING TURNWISE OR FRAMEWISE PREDICTION WITH ACOUSTIC (A) OR ACOUSTIC–LINGUISTIC (A + L) FEATURES: ACCURACY (ACC.), UNWEIGHTED RECALL (REC.), UNWEIGHTED PRECISION (PREC.), AND F1-MEASURE (F1)

model	unit	features	acc.	rec.	prec.	F1
			quadrants + neutral			
BLSTM	turn	A	39.8 %	40.1 %	38.4 %	39.2 %
BLSTM	turn	A+L	41.9 %	41.8 %	41.7 %	**41.7 %**
BLSTM	frame	A	28.0 %	25.3 %	29.5 %	27.2 %
BLSTM	frame	A+L	29.0 %	32.3 %	25.8 %	28.7 %
LSTM	turn	A	40.0 %	38.7 %	36.0 %	37.3 %
LSTM	turn	A+L	41.9 %	41.5 %	37.1 %	39.2 %
LSTM	frame	A	27.8 %	28.6 %	29.6 %	29.1 %
LSTM	frame	A+L	30.4 %	30.0 %	24.7 %	27.1 %
RNN	turn	A	38.0 %	39.8 %	35.4 %	37.5 %
RNN	turn	A+L	39.0 %	41.6 %	37.1 %	39.2 %
RNN	frame	A	28.7 %	24.3 %	25.0 %	24.6 %
RNN	frame	A+L	27.0 %	25.6 %	26.4 %	26.0 %
SVM	turn	A	34.8 %	35.8 %	35.2 %	35.5 %
SVM	turn	A+L	34.8 %	35.9 %	35.0 %	35.4 %
lab	*turn*		*56.8 %*	*55.1 %*	*53.7 %*	*54.3 %*
lab	*frame*		*56.3 %*	*56.9 %*	*54.9 %*	*55.8 %*

TABLE VIII
REGRESSION-(B)LSTM AND RNN PERFORMANCE, SUPPORT VECTOR REGRESSION (SVR) PERFORMANCE, AND AVERAGE LABELER (LAB) CONSISTENCY FOR CLASSIFICATION OF VALENCE AND ACTIVATION (HIGH VERSUS LOW) USING TURNWISE OR FRAMEWISE PREDICTION WITH ACOUSTIC (A) OR ACOUSTIC–LINGUISTIC (A + L) FEATURES: ACCURACY (ACC.), UNWEIGHTED RECALL (REC.), UNWEIGHTED PRECISION (PREC.), AND F1-MEASURE (F1)

model	unit	features	acc.	rec.	prec.	F1
			activation			
BLSTM	turn	A	64.8 %	65.0 %	64.9 %	64.9 %
BLSTM	turn	A+L	64.1 %	64.3 %	64.1 %	64.2 %
BLSTM	frame	A	64.0 %	64.1 %	64.1 %	64.1 %
BLSTM	frame	A+L	65.7 %	65.7 %	65.6 %	**65.6 %**
LSTM	turn	A	59.8 %	60.9 %	61.3 %	61.1 %
LSTM	turn	A+L	60.2 %	60.7 %	60.7 %	60.7 %
LSTM	frame	A	56.4 %	57.2 %	57.4 %	57.3 %
LSTM	frame	A+L	59.1 %	59.9 %	60.1 %	60.0 %
RNN	turn	A	54.6 %	55.1 %	55.2 %	55.2 %
RNN	turn	A+L	55.6 %	56.4 %	56.5 %	56.5 %
RNN	frame	A	53.4 %	55.1 %	56.4 %	55.7 %
RNN	frame	A+L	49.3 %	49.4 %	49.4 %	49.4 %
SVR	turn	A	53.8 %	53.3 %	53.3 %	53.3 %
SVR	turn	A+L	55.5 %	55.2 %	55.8 %	55.2 %
lab	*turn*		*68.6 %*	*70.6 %*	*71.6 %*	*71.1 %*
lab	*frame*		*67.7 %*	*69.4 %*	*70.1 %*	*69.8 %*
			valence			
BLSTM	turn	A	56.5 %	58.0 %	58.3 %	58.1 %
BLSTM	turn	A+L	60.0 %	61.1 %	61.4 %	61.3 %
BLSTM	frame	A	65.8 %	64.0 %	64.7 %	64.3 %
BLSTM	frame	A+L	72.8 %	72.2 %	72.1 %	**72.2 %**
LSTM	turn	A	61.0 %	62.5 %	62.9 %	62.7 %
LSTM	turn	A+L	58.8 %	60.3 %	60.9 %	60.6 %
LSTM	frame	A	55.9 %	57.4 %	57.4 %	57.4 %
LSTM	frame	A+L	63.6 %	57.7 %	67.3 %	62.1 %
RNN	turn	A	58.8 %	60.3 %	60.8 %	60.5 %
RNN	turn	A+L	62.9 %	64.2 %	64.8 %	64.5 %
RNN	frame	A	60.9 %	63.6 %	64.3 %	63.9 %
RNN	frame	A+L	57.5 %	62.0 %	66.0 %	63.9 %
SVR	turn	A	53.1 %	55.0 %	55.6 %	55.3 %
SVR	turn	A+L	56.0 %	57.5 %	58.0 %	57.8 %
lab	*turn*		*88.6 %*	*88.4 %*	*88.6 %*	*88.6 %*
lab	*frame*		*86.0 %*	*85.8 %*	*85.6 %*	*85.7 %*

The best F1-measure for valence (72.2%) is notably below the average "performance" or consensus of a human labeler (85.7%). However, the best recognition result for activation (68.9%) is only 2.2% below the inter-human labeling consistency (71.1%). For the five-class task the performance gap between the best classifier and human labelers is 8.6% (see Table V).

In what follows, we will analyze the results in Tables IV–IX with respect to six different aspects: the number of emotion classes, the difference between regression and discriminative training, the gain of LSTM context modeling, the benefit of including bidirectional context, the difference between turnwise and framewise classification, and the integration of linguistic features.

1) Four Quadrants Versus Five Classes: The best F1-measure for quadrant classification can be obtained when using a discriminative BLSTM for turnwise prediction with acoustic features (51.3%, see Table VI). However, additionally modeling the "neutral" state can lead to a comparable prediction performance (47.2%, see Table V). Interestingly, for the five-class task framewise regression prevails. Obviously, the higher number of class borders a discriminative classifier has to face in the five-class experiment downgrades performance significantly. As can be seen in Table V, a BLSTM network modeling all five classes profits from frame by frame modeling of the fineness of emotional dynamics via regression. Tables X and XI show typical confusions when distinguishing four and five classes, respectively. In both cases, the best prediction quality can be obtained for quadrant IV (*angry/anxious*). Table XI points out that, due to the non-prototypicality of emotions in the SAL corpus, almost all quadrants are most frequently confused with the neutral state. An impression of the prediction quality for more prototypical utterances (or utterances with emotions of higher intensity) can be obtained when masking the last column and the last line of Table XI: quadrant–quadrant confusions obviously occur less frequent than quadrant–neutral confusions. Another interesting aspect is the effect of emotional intensity—and thus indirectly prototypicality—of the test set on the obtained recognition performance: when using the Regression-BLSTM for framewise prediction with acoustic and linguistic features (trained on *all*

TABLE IX
DISCRIMINATIVE-(B)LSTM AND RNN PERFORMANCE, SUPPORT VECTOR MACHINE (SVM) PERFORMANCE, AND AVERAGE LABELER (LAB) CONSISTENCY FOR CLASSIFICATION OF VALENCE AND ACTIVATION (HIGH VERSUS LOW) USING TURNWISE OR FRAMEWISE PREDICTION WITH ACOUSTIC (A) OR ACOUSTIC–LINGUISTIC (A + L) FEATURES: ACCURACY (ACC.), UNWEIGHTED RECALL (REC.), UNWEIGHTED PRECISION (PREC.), AND F1-MEASURE (F1)

model	unit	features	acc.	rec.	prec.	F1
			activation			
BLSTM	turn	A	68.3 %	68.9 %	68.8 %	**68.9 %**
BLSTM	turn	A+L	66.4 %	66.5 %	66.4 %	66.4 %
BLSTM	frame	A	62.8 %	63.6 %	64.0 %	63.8 %
BLSTM	frame	A+L	58.0 %	57.9 %	57.8 %	57.9 %
LSTM	turn	A	63.4 %	64.8 %	65.6 %	65.2 %
LSTM	turn	A+L	65.3 %	66.2 %	66.5 %	66.4 %
LSTM	frame	A	50.0 %	50.8 %	50.8 %	50.8 %
LSTM	frame	A+L	56.3 %	56.8 %	56.9 %	56.9 %
RNN	turn	A	61.7 %	63.0 %	63.8 %	63.4 %
RNN	turn	A+L	61.5 %	62.9 %	63.7 %	63.3 %
RNN	frame	A	50.6 %	52.7 %	53.8 %	53.3 %
RNN	frame	A+L	54.4 %	55.2 %	55.4 %	55.3 %
SVM	turn	A	55.8 %	56.7 %	56.8 %	56.8 %
SVM	turn	A+L	54.4 %	55.2 %	55.3 %	55.3 %
lab	*turn*		*68.6 %*	*70.6 %*	*71.6 %*	*71.1 %*
lab	*frame*		*67.7 %*	*69.4 %*	*70.1 %*	*69.8 %*
			valence			
BLSTM	turn	A	63.7 %	64.6 %	64.7 %	64.7 %
BLSTM	turn	A+L	71.2 %	71.8 %	71.7 %	**71.7 %**
BLSTM	frame	A	63.8 %	65.1 %	64.8 %	65.0 %
BLSTM	frame	A+L	55.0 %	58.4 %	59.7 %	59.0 %
LSTM	turn	A	56.4 %	59.4 %	63.4 %	61.3 %
LSTM	turn	A+L	66.8 %	68.5 %	70.1 %	69.3 %
LSTM	frame	A	65.3 %	66.3 %	65.9 %	66.1 %
LSTM	frame	A+L	58.3 %	56.1 %	56.6 %	56.4 %
RNN	turn	A	67.5 %	67.9 %	67.8 %	67.9 %
RNN	turn	A+L	69.5 %	70.5 %	70.6 %	70.5 %
RNN	frame	A	57.5 %	60.3 %	61.0 %	60.6 %
RNN	frame	A+L	64.2 %	64.6 %	64.2 %	64.4 %
SVM	turn	A	61.4 %	63.5 %	65.7 %	64.6 %
SVM	turn	A+L	59.3 %	61.4 %	62.9 %	62.1 %
lab	*turn*		*88.6 %*	*88.4 %*	*88.6 %*	*88.6 %*
lab	*frame*		*86.0 %*	*85.8 %*	*85.6 %*	*85.7 %*

TABLE X
CONFUSION MATRIX FOR THE BEST QUADRANT CLASSIFICATION SETTING (DISCRIMINATIVE BLSTM FOR TURNWISE PREDICTION WITH ACOUSTIC FEATURES ONLY); ROWS: GROUND TRUTH; COLUMNS: PREDICTIONS (WHITE TO BLACK RESEMBLES 0–100 %)

%	I	II	III	IV
I	39	31	9	21
II	9	54	12	25
III	4	27	47	22
IV	3	21	9	67

training data and characterized by the five-class confusion matrix in Table XI), while evaluating only those utterances that are *not* annotated as "neutral," the resulting quadrant prediction F1-measure is 58.2%. On the other hand, when evaluating only those turns that are annotated as "neutral," the F1-measure for quadrant prediction is as low as 34.3%. For very "intense" test utterances that are labeled as having an absolute value of activation and valence that is higher than 0.5, the obtained quadrant prediction F1-measure is 85.1%.

2) Regression Versus Discriminative Training: For almost every experimental setting we can observe that discriminative

TABLE XI
CONFUSION MATRIX FOR THE BEST "QUADRANTS + NEUTRAL" (N) CLASSIFICATION SETTING (REGRESSION BLSTM FOR FRAMEWISE PREDICTION WITH ACOUSTIC AND LINGUISTIC FEATURES); ROWS: GROUND TRUTH; COLUMNS: PREDICTIONS (WHITE TO BLACK RESEMBLES 0–100 %)

%	I	II	III	IV	N
I	40	13	6	4	37
II	25	40	3	8	24
III	12	1	48	14	25
IV	2	9	1	80	8
N	22	11	10	16	41

training prevails for turnwise recognition while regression prevails for framewise recognition. Complete turns that are characterized by statistical functionals of features can be distinguished better with a discriminative technique. On the other hand, when predicting a class frame by frame the network fails to model "label jumps" when discriminatively trained on the discrete labels. For framewise prediction, modeling the smooth progression of valence and activation is necessary before mapping the output activations to quadrants.

3) LSTM Context Modeling Versus RNN and SVM: Both, for framewise but also for turnwise prediction the LSTM architecture outperforms a conventional RNN in most cases. The major reason for this is the *vanishing gradient problem* (see Section III) which limits the amount of context a recurrent neural network can access. Using no contextual information at all leads to comparatively low performance as can be seen in the SVR and SVM experiments, justifying the higher computational cost of the LSTM approach.

4) Unidirectional Versus Bidirectional Context: Independent of the classification task, bidirectional context mostly prevails over unidirectional context. Both, regression and discriminative BLSTM networks outperform all other models (LSTM, RNN, SVR, and SVM) for the discrimination of five, four, and two classes (numbers in bold face in Tables IV–IX).

5) Turnwise Versus Framewise Classification: As already mentioned, turnwise prediction can successfully be combined with discriminative learning, while framewise emotion recognition is rather suited for predictors based on regression. For both strategies, modeling contextual information is essential. When additionally modeling "neutrality," the best result can be obtained with framewise prediction (see Table V). Note that the amount of contextual information a BLSTM network models is a lot more flexible when framewise prediction is applied, since the temporal granularity is higher than it is for turnwise recognition. This can be seen as the major reason why framewise recognition outperforms turnwise prediction if regression-BLSTM networks are used.

6) Acoustic Features Versus Combined Acoustic and Linguistic Features: Comparing Tables IV and VI, one can assert that the regression-LSTM seems to profit more from the inclusion of linguistic features. In some cases the quadrant prediction performance of the discriminative classifier is even degraded when adding keyword features. Obviously, the presence of single keywords is not discriminative enough in this case. Linguistic features are rather suited for modeling tendencies within a continuous scale for valence and activation. When modeling

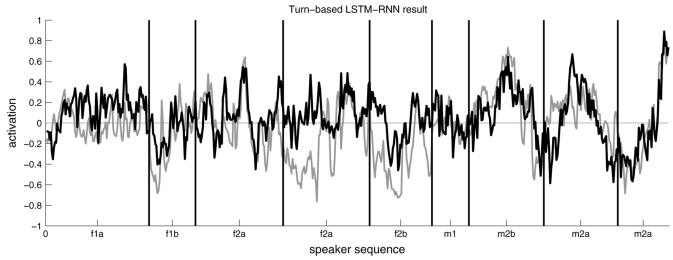

Fig. 7. Prediction of activation (black) using a regression-LSTM and ground truth (gray) over all turns of the test set (only acoustic features used).

"neutrality" as a fifth class, also the *discriminative* BLSTM profits from linguistic features (while this is not the case for the discriminative *four-class* task). This supports the finding that a performance gain through keyword features presumes a certain level of granularity of the prediction targets.

As an example for emotion recognition using regression, Fig. 7 shows the turnwise activation predictions of a regression-LSTM before the output activations are mapped onto quadrants. Prediction and ground truth are correlated with a correlation coefficient of 0.56, leading to an F1-measure of 61.1% (see Table VIII) when distinguishing positive and negative activation for every speech turn.

VII. Conclusion

In this paper, we introduced a novel technique for the estimation of the quadrant in a two-dimensional emotional space spanned by the dimensions *valence* and *activation*, as it is needed for the SAL—an emotionally sensitive virtual agent developed within the SEMAINE project. In contrast to many other works that report recognition results for the static classification of acted speech turns representing *emotional prototypes*, our contribution can be seen as a realistic evaluation of recognition accuracy under real-life conditions, where non-prototypical speech has to be classified using powerful techniques of dynamic speech modeling. Our approach combines acoustic features obtained by our openEAR online feature extractor with binary linguistic features produced by a tandem LSTM-DBN, which are then classified by a long short-term memory recurrent neural net. The LSTM architecture allows for the modeling of long-range contextual information and enables a new technique of incremental affect recognition that does not require the computation of statistical functionals of features but captures the temporal evolution indirectly through LSTM memory cells. As an alternative for regression-based quadrant prediction, we designed a discriminatively trained LSTM network which explicitly learns to distinguish quadrants of the emotional space. The design of our proposed AER system is based on a series of findings documented in earlier works: the benefit of including linguistic features for speech based emotion recognition [14], the enhancement of keyword spotting performance through the incorporation of LSTM phoneme prediction features [41], the importance of modeling temporal long-range dependencies in emotion recognition [26], and the potential of discriminative learning for quadrant prediction [85]. The prediction quality of our system was shown to be comparable to the degree of consistency between different human labelers.

One short-coming of our system is the fact that *bidirectional* context cannot be used in a causal online emotion recognition system. However, since we observed improved results for bidirectional LSTM networks, the investigation of the potential of BLSTM-RNN for online recognition is promising. For future experiments, 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. A further drawback of the introduced system is its complexity. However, provided that only *unidirectional* context is used, our system can still operate in real-time. The training of the complete system as used in this paper can be completed within one day, but will take longer as soon as larger training databases are used. Another problem—implied by the recognition task—is that our classification system has to deal with a high amount of ambiguous speech turns which are near the class borders in the valence-activation space. This leads to high error rates for non-prototypical speech segments that are difficult to model when using discrete classes. A possible solution is to continuously model emotion via regression while abstaining from mapping the regression output onto quadrants. Yet, those continuous values are difficult to use for the dialogue management system of an emotion-sensitive virtual agent which will have to use thresholds or any other kind of discretization before selecting adequate system responses. As far as AER performance evaluation is concerned, a possible solution is to increase the granularity of emotional space discretization (e.g., by defining nine instead of four regions in the emotional space) while at the same time tolerating confusions between neighboring regions, as done in [26], for example. Even though "wrong" assignments of ambiguous speech turns are not necessarily critical for the quality or adequateness of a virtual agent's responses (even humans can interpret such utterances differently), further research will be necessary in this area.

Future works will focus on investigating the benefit of including further feature types, such as vision features used in [14] or [88], into a time-continuous context sensitive emotion recognition framework. For this purpose it would be interesting to examine the potential of hybrid fusion techniques such as asynchronous hidden Markov models [89] or multidimensional dynamic time warping [90] as alternatives to late and early fusion. Also the LSTM architecture and parameterization could be optimized by including more hidden layers or using different layer sizes. Furthermore it would be interesting to examine the potential of multi-task learning, i.e., learning the phonemes and the affective state simultaneously. In addition to the mentioned approaches for future improvements, there will be a lot more aspects to consider before emotion-sensitive systems can show a degree of naturalness that is comparable to humans. Yet, 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, the principle of incremental speech processing and the integration of long-range context information can be seen as two further steps towards making virtual agents more human-like.

REFERENCES

[1] M. T. Vo and A. Waibel, "Multimodal human–computer interaction," in *Proc. ISSD'93*, Waseda, Japan, 1993.

[2] S. Furui, "Toward the ultimate synthesis/recognition," *Proc. Nat. Acad. Sci. USA*, vol. 92, no. 22, pp. 10 040–10 045, 1995.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human–computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Feb. 2001.

[4] R. W. Picard, "Toward agents that recognize emotion," in *Actes Proc. IMAGINA*, 1998, pp. 153–165.

[5] E. Shriberg, "How peoply really talk and why engineers should care," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 1781–1784.

[6] Z. Zeng, M. Pantic, G. I. Rosiman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[7] R. Picard, *Affective Computing*. Cambridge, MA: MIT Press, 1997.

[8] S. Casale, A. Russo, G. Scebba, and S. Serrano, "Speech emotion classification using machine learning algorithms," in *Proc. IEEE Int. Conf. Semantic Comput.*, 2008, pp. 158–165.

[9] B. Schuller, M. Wimmer, L. Mösenlechner, C. Kern, D. Arsic, and G. Rigoll, "Brute-forcing hierarchical functionals for paralinguistics: A waste of feature space?," in *Proc. ICASSP'08*, Las Vegas, NV, 2008, pp. 4501–4504.

[10] M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[11] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Netw.*, vol. 18, no. 4, pp. 407–422, 2005.

[12] ***[AUTHOR: Please provide page range]***V. Petrushin, "Emotion in speech: Recognition and application to call centers," *Artif. Neural Netw Eng. (ANNIE)*, 1999.

[13] B. Schuller, R. Müller, B. Hörnler, A. Hoethker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," in *Proc. Int. Conf. Multimodal Interfaces, ACM SIGHI*, Nagoya, Japan, 2007, pp. 30–37.

[14] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? recognising natural interest by extensive audiovisual integration for real-life application," *Image Vis. Comput. J. (IMAVIS), Special Iss. Vis. Multimodal Anal. Human Spontaneous Behavior*, vol. 27, no. 12, pp. 1760–1774, 2009.

[15] A. Batliner, S. Steidl, and E. Nöth, "Releasing a thoroughly annotated and processed spontaneous emotional database: The FAU Aibo Emotion Corpus," in *Proc. Satellite Workshop of LREC 2008 Corpora Res. Emotion and Affect*, L. Devillers, J. C. Martin, R. Cowie, E. Douglas-Cowie, and A. Batliner, Eds., 2008, pp. 28–31.

[16] S. Steininger, F. Schiel, O. Dioubina, and S. Raubold, "Development of user-state conventions for the multimodal corpus in smartkom," in *Proc. Workshop Multimodal Resources Multimodal Syst. Eval.*, Las Palmas, Spain, 2002, pp. 33–37.

[17] B. Schuller, G. Rigoll, S. Can, and H. Feussner, "Emotion sensitive speech control for human-robot interaction in minimal invasive surgery," in *Proc. 17th Int. Symp. Robot Human Interactive Commun. RO-MAN'08*, Munich, Germany, 2008, pp. 453–458.

[18] J. Hansen and S. Bou-Ghazale, "Getting started with SUSAS: A speech under simulated and actual stress database," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 1743–1746.

[19] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Proc. ICME*, Amsterdam, The Netherlands, 2005.

[20] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech'05*, Lisbon, Portugal, 2005, pp. 1517–1520.

[21] I. S. Engberg, A. V. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a Danish emotional speech database," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. 1695–1698.

[22] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 Audiovisual Emotion Database," in *Proc. IEEE Workshop Multimedia Database Management*, Atlanta, Georgia, 2006.

[23] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 emotion challenge," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 312–315.

[24] S. Steidl, B. Schuller, A. Batliner, and D. Seppi, "The hinterland of emotions: Facing the open-microphone challenge," in *Proc. ACII*, Amsterdam, The Netherlands, 2009, pp. 690–697.

[25] M. Schröder, R. Cowie, D. Heylen, M. Pantic, C. Pelachaud, and B. Schuller, "Towards responsive sensitive artificial listeners," in *Proc. 4th Int. Workshop Human-Comput. Convers.*, Bellagio, Italy, 2008.

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

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, no. 5–6, pp. 602–610, Jun. 2005.

[29] M. Wöllmer, F. Eyben, J. Keshet, A. Graves, B. Schuller, and G. Rigoll, "Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks," in *Proc. ICASSP*, Taipei, Taiwan, 2009, pp. 3949–3952.

[30] M. Wöllmer, F. Eyben, B. Schuller, Y. Sun, T. Moosmayr, and N. Nguyen-Thien, "Robust in-car spelling recognition—A tandem BLSTM-HMM approach," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 2507–2510.

[31] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural Comput. Applicat.*, vol. 9, pp. 290–296, 2000.

[32] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[33] B. Schuller, S. Reiter, and G. Rigoll, "Evolutionary feature generation in speech emotion recognition," in *Proc. ICME'06*, Toronto, ON, Canada, 2006, pp. 5–8.

[34] M. Streit, A. Batliner, and T. Portele, "Emotions analysis and emotion-handling subdialogues," in *SmartKom: Foundations of Multimodal Dialogue Systems*, W. Wahlster, Ed. Berlin, Germany: Springer, 2006, pp. 317–332.

[35] S. Steidl, in *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*, Berlin, Germany, 2009, Logos Verlag.

[36] B. Schuller and G. Rigoll, "Timing levels in segment-based speech emotion recognition," in *Proc. Interspeech'06*, Pittsburgh, PA, 2006, pp. 1818–1821, ISCA.

[37] B. Schuller, B. Vlasenko, R. Minguez, G. Rigoll, and A. Wendemuth, "Comparing one and two-stage acoustic modeling in the recognition of emotion in speech," in *Proc. ASRU'07*, Kyoto, Japan, 2007, pp. 596–600.

[38] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. ICASSP'03*, Hong Kong, China, 2003, pp. 1–4.

[39] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. turnlevel: Emotion recognition from speech considering static and dynamic processing," in *Proc. ACII'07, Lisbon, Portugal*, A. Paiva, Ed., Heidelberg, Germany, 2007, vol. LNCS 4738, pp. 139–147, Springer Berlin.

[40] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[41] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll, "A tandem BLSTM-DBN architecture for keyword spotting with enhanced context modeling," in *Proc. NOLISP*, Vic, Spain, 2009.

[42] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J. C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, "The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data," in *Affective Computing and Intelligent Interaction*. Berlin/Heidelberg, Germany: Springer, 2007, vol. 4738/2007, pp. 488–500 [Online]. Available: http://emotionresearch.net/download/pilot-db/

[43] C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 1983–1986.

[44] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "FEELtrace: An instrument for recording perceived emotion in real time," in *Proc. ISCA Workshop Speech Emotion*, 2000, pp. 19–24.

[45] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME)*, Hannover, Germany, 2008, pp. 865–868.

[46] D. Seppi, A. Batliner, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and V. Aharonson, "Patterns, prototypes, performance: Classifying emotional user states," in *Proc. Interspeech*, Brisbane, Australia, 2008, pp. 601–604.
[47] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR—Introducing the Munich open-source emotion and affect recognition toolkit," in *Proc. ACII*, Amsterdam, The Netherlands, 2009, pp. 576–581.
[48] M. Schröder, L. Devillers, K. Karpouzis, J.-C. Martin, C. Pelachaud, C. Peter, H. Pirker, B. Schuller, J. Tao, and I. Wilson, "What should a generic emotion markup language be able to represent?," in *Affective Computing and Intelligent Interaction*, A. Paiva, R. Prada, and R. W. Picard, Eds. Berlin-Heidelberg, Germany: Springer, 2007, pp. 440–451.
[49] P. Baggia, F. Burkhardt, C. Pelachaud, C. Peter, B. Schuller, I. Wilson, and E. Zovato, "Elements of an EmotionML 1.0," in *W3C Incubator Group Report*, M. Schröder, Ed., 2008, W3C.
[50] B. Schuller and G. Rigoll, "Recognising interest in conversational speech-comparing bag of frames and supra-segmental features," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 1999–2002.
[51] A. Quattoni, S. Wang, L. P. Morency, M. Collins, and T. Darrell, "Hidden conditional random fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1848–1853, Oct. 2007.
[52] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "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. Piscataway, NJ: IEEE Press, 2001.
[53] A. Graves, S. Fernandez, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *Proc. ICANN*, Warsaw, Poland, 2005, vol. 18, pp. 602–610.
[54] S. Fernandez, A. Graves, and J. Schmidhuber, "Phoneme Recognition in TIMIT With BLSTM-CTC," IDSIA, Tech. Rep., 2008.
[55] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Univ. of Waikato, Hamilton, New Zealand, 1999.
[56] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.
[57] S. Arunachalam, D. Gould, E. Anderson, D. Byrd, and S. Narayanan, "Politeness and frustration language in child-machine interactions," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 2675–2678.
[58] Z. J. Chuang and C. H. Wu, "Emotion recognition using acoustic features and textual content," in *Proc. ICME*, 2004, pp. 53–56.
[59] K. Dupuis and K. Pichora-Fuller, "Use of lexical and affective prosodic cues to emotion by younger and older adults," in *Proc. Interspeech*, Antwerp, Belgium, 2007, pp. 2237–2240.
[60] C. Elliott, "The affective reasoner: A process model of emotions in a multi-agent system," Ph.D. dissertation, Northwestern Univ., Evanston, IL, 1992.
[61] R. Cowie, E. Douglas-Cowie, B. Apolloni, J. Taylor, A. Romano, and W. Fellenz, "What a neural net needs to know about emotion words," *Comput. Intell. Applicat.*, pp. 109–114, 1999, N. Mastorakis, Ed..
[62] D. Litman and K. Forbes, "Recognizing emotions from student speech in tutoring dialogues," in *Proc. ASRU*, Virgin Islands, 2003, pp. 25–30.
[63] X. Zhe and A. Boucouvalas, "Text-to-emotion engine for real time internet communication," in *Proc. Int. Symp. Commun. Syst., Netw., DSPs*, 2002, pp. 164–168, Staffordshire Univ., Stoke-on-Trent, U.K.
[64] B. Goertzel, K. Silverman, C. Hartley, S. Bugaj, and M. Ross, "The baby webmind project," in *Proc. Annu. Conf. Soc. Study Artif. Intell. Simul. Behav. (AISB)*, 2000.
[65] T. Wu, F. Khan, T. Fisher, L. Shuler, and W. Pottenger, "Posting act tagging using transformation-based learning," in *Foundations of Data Mining and Knowledge Discovery*, T. Y. Lin, S. Ohsuga, C. J. Liau, X. Hu, and S. Tsumoto, Eds., 2005, pp. 319–331.
[66] H. Liu, H. Liebermann, and T. Selker, "A model of textual affect sensing using real-world knowledge," in *Proc. IUI*, 2003, pp. 125–132.
[67] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Proc. ICASSP*, 2004, pp. 577–580.
[68] J. Breese and G. Ball, "Modeling emotional state and personality for conversational agents," Microsoft, Tech. Rep., 1998.
[69] G. Rigoll, R. Müller, and B. Schuller, "Speech emotion recognition exploiting acoustic and linguistic information sources," in *Proc. SPECOM*, Patras, Greece, 2005, pp. 61–67.
[70] T. S. Polzin and A. Waibel, "Emotion-sensitive human–computer interfaces," in *Proc. ISCA ITRW Speech Emotion*, 2000, pp. 201–206.
[71] J. Ang, R. Dhillon, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human–computer dialog," in *Proc. Interspeech*, Denver, CO, 2002, pp. 2037–2040.
[72] C. M. Lee, S. Narayanan, and R. Pieraccini, "Combining acoustic and language information for emotion recognition," in *Proc. ICSLP*, 2002, pp. 873–376.
[73] L. Devillers, L. Lamel, and I. Vasilescu, "Emotion detection in task-oriented spoken dialogs," in *Proc. ICME*, Baltimore, MD, 2003.
[74] B. Schuller, R. Müller, M. Lang, and G. Rigoll, "Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensemble," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 805–808.
[75] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "Combining efforts for improving automatic classification of emotional user states," in *Proc. IS-LTC*, Ljubliana, Slovenia, 2006, pp. 240–245.
[76] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. ECML*, Chemniz, Germany, 1998, pp. 137–142.
[77] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Emotion recognition from speech: Putting asr in the loop," in *Proc. ICASSP*, Taipei, Taiwan, 2009, pp. 4585–4588.
[78] B. Schuller, J. Schenk, and G. Rigoll, ""The Godfather" vs. "chaos": Comparing linguistic analysis based on online knowledge sources and bags-of-n-grams for movie review valence estimation," in *Proc. ICDAR*, Barcelona, Spain, 2009, pp. 858–862.
[79] Y. Wang and I. H. Witten, "Modeling for optimal probability prediction," in *Proc. 19th Int. Conf. Mach. Learn.*, Sydney, Australia, 2002, pp. 650–657.
[80] M. Wöllmer, F. Eyben, B. Schuller, and G. Rigoll, "Robust vocabulary independent keyword spotting with graphical models," in *Proc. ASRU*, Merano, Italy, 2009, pp. 349–353.
[81] J. A. Bilmes, "Graphical models and automatic speech recognition," in *Mathematical Foundations of Speech and Language Processing*, R. Rosenfeld, M. Ostendorf, S. Khudanpur, and M. Johnson, Eds. New York: Springer Verlag, 2003, pp. 191–246.
[82] F. V. Jensen, *An Introduction to Bayesian Networks*. New York: Springer, 1996.
[83] J. Bilmes and G. Zweig, "The graphical models toolkit: An open source software system for speech and time-series processing," in *Proc. ICASSP*, 2002, pp. 3916–3919.
[84] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM," 1993.
[85] M. Wöllmer, F. Eyben, B. Schuller, E. Douglas-Cowie, and R. Cowie, "Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 1595–1598.
[86] M. Riedmiller and H. Braun, "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in *Proc. IEEE Int. Conf. Neural Netw.*, 1993, pp. 586–591.
[87] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *Proc. ICASSP*, Honolulu, HI, 2007, pp. 1085–1088.
[88] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: Face, body gesture, speech," in *Affect and Emotion in Human-Computer Interaction*, C. Peter and R. Beale, Eds. New York: Springer, 2008, vol. 4868, lNCS.
[89] S. Bengio, "An asynchronous hidden Markov model for audio-visual speech recognition," Advances in NIPS 15 2003.
[90] M. Wöllmer, M. Al-Hames, F. Eyben, B. Schuller, and G. Rigoll, "A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams," *Neurocomputing*, vol. 73, pp. 366–380, 2009.

Martin Wöllmer (M'09) received the diploma in electrical engineering and information technology from the Technische Universität München (TUM), Munich, Germany.

He works as a Researcher funded by the European Community's Seventh Framework Program project SEMAINE (FP7/2007–2013) at TUM, where his current research and teaching activity includes the subject areas of pattern recognition and speech processing. His focus lies on multimodal data fusion, automatic recognition of emotionally colored and noisy speech, and speech feature enhancement. Publications of his in various

journals and conference proceedings cover novel and robust modeling architectures for speech and emotion recognition such as switching linear dynamic models or long short-term memory recurrent neural nets.

Mr. Wöllmer is a Reviewer for the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING and the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING.

Björn Schuller (M'04) received the diploma and Ph.D. degrees in electrical engineering and information technology from Technische Universität München (TUM), Munich, Germany.

He is currently a Lecturer in pattern recognition at TUM. He authored more than 120 publications in books, journals, and peer-reviewed conference proceedings in this field. Best known are his works advancing audiovisual processing in the areas of affective computing and multimedia retrieval.

Dr. Schuller serves as a member of the steering committee of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, as associate editor and reviewer for several scientific journals, including the IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, *Elsevier Signal Processing*, and *Speech Communication*, and as invited speaker, session organizer, and chairman, and program committee member of numerous international conferences. Current project steering board activities include SEMAINE funded by the European Community and further projects with companies as BMW, Continental, Daimler, Siemens, Toyota, and VDO. He is invited expert in the W3C Emotion and Emotion Markup Language Incubator Groups, and elected member of HUMAINE Association Executive Committee.

Florian Eyben (M'09) received the diploma in information technology from the Technische Universität München (TUM), Munich, Germany.

He works on a research grant as part of the European Community's Seventh Framework Program project SEMAINE (FP7/2007–2013)—The Sensitive Artificial Listener project—within the Institute for Human–Machine Communication at TUM. Teaching activities of his comprise pattern recognition and speech and language processing. His research interests include large-scale hierarchical audio feature extraction and evaluation, automatic emotion recognition from the speech signal, recognition of non-linguistic vocalizations, automatic continuous large-vocabulary speech recognition, statistical and context-dependent language models, and music information retrieval. He has several publications in various journals and conference proceedings covering many of his areas of research.

Gerhard Rigoll (M'86–SM'98) received the diploma in technical cybernetics in 1982, the Ph.D. degree for his work in the field of automatic speech recognition in 1986, and the habilitation in the field of speech synthesis in 1991, all from the University of Stuttgart, Stuttgart, Germany.

He was with the Fraunhofer-Institute Stuttgart, Speech Plus in Mountain View, CA, and Digital Equipment in Maynard, MA, spent a Post-Doctoral Fellowship at IBM T. J. Watson Research Center, Yorktown Heights, NY, headed a research group at Fraunhofer-Institute Stuttgart, and spent a two year's research stay at NTT Human Interface Laboratories in Tokyo, Japan, in 1986, in the area of neuro-computing, speech recognition and pattern recognition until he was appointed as a Full Professor of computer science at Gerhard-Mercator-University, Duisburg, Germany, 1993 and of Human–Machine Communication at the Technische Universität München (TUM), Munich, Germany, in 2002. He authored and coauthored more than 250 publications in the field of signal processing and pattern recognition. Most of his work deals with automatic speech recognition, where he is particularly concerned with classifier optimization. He also maintains active research programs in vision-based pattern recognition.