Computational assessment of interest in speech - facing the real-life challenge

Martin Wöllmer, Felix Weninger, Florian Eyben, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Wöllmer, Martin, Felix Weninger, Florian Eyben, and Björn Schuller. 2011. "Computational assessment of interest in speech - facing the real-life challenge." *KI- Künstliche Intelligenz* 25 (3): 225. https://doi.org/10.1007/s13218-011-0108-9.

Computational Assessment of Interest in Speech—Facing the Real-Life Challenge

Martin Wöllmer · Felix Weninger · Florian Eyben · Björn Schuller

Abstract Automatic detection of a speaker's level of interest is of high relevance for many applications, such as automatic customer care, tutoring systems, or affective agents. However, as the latest Interspeech 2010 Paralinguistic Challenge has shown, reliable estimation of non-prototypical natural interest in spontaneous conversations independent of the subject still remains a challenge. In this article, we introduce a fully automatic combination of brute-forced acoustic features, linguistic analysis, and non-linguistic vocalizations, exploiting cross-entity information in an early feature fusion. Linguistic information is based on speech recognition by a multi-stream approach fusing context-sensitive phoneme predictions and standard acoustic features. We provide subject-independent results for interest assessment using Bidirectional Long Short-Term Memory networks on the official Challenge task and show that our proposed system leads to the best recognition accuracies that have ever been reported for this task. The according TUM AVIC corpus consists of highly spontaneous speech from face-to-face commercial presentations. The techniques presented in this article are also used in the SEMAINE system, which features an emotion sensitive embodied conversational agent.

Keywords Affective computing · Interest recognition · Recurrent neural networks · Long short-term memory

1 Introduction

Automatically extracting information on interest or disinterest of users possesses great potential for general Human-Computer Interaction [16, 26, 37] and many applications, including sales and advertisement systems, virtual guides, or conversational agents like the SEMAINE system [19]. Recent studies on interest recognition have primarily focused on use-cases such as topic switching in infotainment or customer service systems [20], meeting analysis [3, 13, 27], or (children's) tutoring systems [14]. In this context, the organizers of the Interspeech 2010 Paralinguistic Challenge [23] defined an interest recognition task with unified system training and test conditions in order to make the recognition approaches developed by different researchers easily comparable. In the *Affect Sub-Challenge*, the task is to automatically predict a user's level of interest from the speech signal applying a pre-defined acoustic feature set and (optionally) linguistic information. The corpus used for training and evaluation is the Audiovisual Interest Corpus recorded at the Technische Universität München ("TUM AVIC") [20]. It features highly spontaneous speech from face-to-face commercial presentations and reflects the conditions a real-life interest recognition system has to face. The challenge task—predicting a speaker's level of interest in ordinal representation by suited regression techniques—deliberately avoids hard decisions as it is well known that human affect is also continuous and cannot be sufficiently described by a limited set of categories. In this article, we present our recent research on affective state recognition by introducing a fully automatic interest recognition system as it is applied in the SEMAINE system [19]—a virtual conversational agent tailored to emotionally sensitive topic-independent human-machine interaction. In contrast to the baseline Paralinguistic Challenge recognition system that has been applied and evaluated in [23] and is based on acoustic features processed via unpruned REP-Trees, our proposed system also makes use of linguistic information obtained by automatic speech recognition (ASR) and exploits a self-learned amount of

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

contextual information. We apply Bidirectional Long Short-Term Memory (BLSTM) recurrent neural networks which have shown impressive affect recognition performance gains with respect to standard techniques such as Support Vector Regression or Hidden Markov Models (HMM) [35, 36]. Furthermore, since ASR in conversational speech scenarios tends to be more challenging than, e.g., the recognition of read speech, we build on our recent work on robust ASR by applying a multi-stream BLSTM-HMM system [34] for extracting linguistic information. The multi-stream model is composed of a BLSTM network for context-sensitive phoneme prediction and an HMM that uses both, BLSTM-based phoneme prediction features and conventional Mel-Frequency Cepstral Coefficient (MFCC) features as observations. The combined acoustic-linguistic information for interest recognition is represented in a joint feature vector via early fusion.

In Sect. 2, we provide details on the TUM AVIC corpus and introduce the recognition task. Next, in Sect. 3, we motivate the use of Long Short-Term Memory and give a brief overview over the BLSTM architecture. Section 4 contains information about acoustic and linguistic feature extraction and about the employed feature selection technique. In Sect. 5, we outline experiments and results before drawing conclusions in Sect. 6.

2 The TUM AVIC Corpus

The experiments outlined in Sect. 5 are based on the “TUM AVIC” corpus [20] which had also been used for the Affect Sub-Challenge of the Interspeech 2010 Paralinguistic Challenge [23]. In the scenario setup, an experimenter and a subject are sitting on opposite sides of a desk. The experimenter plays the role of a product presenter and leads the subject through a commercial (car) presentation. The subject’s role is to listen to explanations and topic presentations of the experimenter, ask several questions of her/his interest, and actively interact with the experimenter considering his/her interest in the addressed topics. To capture a large variety of interest levels, the subject was explicitly asked not to worry about being polite to the experimenter, e.g., by always showing a certain level of ‘polite’ attention. Instead the participants were encouraged to honestly express interest or disinterest, depending on the content of the presentation. Visual and speech data was recorded by a camera and two microphones, one headset and one far-field microphone. In conformance with the Interspeech 2010 Paralinguistic Challenge, we exclusively use data recorded by the lapel microphone (44.1 kHz, 16 bit).

21 subjects took part in the recordings, three of them Asian, the remaining European. The language throughout experiments is English, and all subjects are non-native, yet very experienced English speakers. The mean age of the participants is 29.9 years and the total recording time is 10.4 h. To acquire reliable labels of a subject’s ‘level of interest’ (LOI), the entire video material was segmented into speaker and sub-speaker-turns and subsequently labeled by four male annotators, independently from each other. The annotators were undergraduate students of psychology. The intention was to annotate observed interest in the common sense. A speaker-turn is defined as continuous speech segment produced solely by one speaker—back channel interjections (“mhm”, etc.) are ignored, i.e., every time there is a speaker change, a new speaker turn begins. This is in accordance with the common understanding of the term ‘turn-taking’. Speaker-turns thus can contain multiple and especially long sentences. In order to provide level of interest analysis on a finer time scale, the speaker turns were further segmented at grammatical phrase boundaries: a turn lasting longer than two seconds is split by punctuation and syntactical and grammatical rules, until each segment is shorter than two seconds. These resulting segments are referred to as sub-speaker-turns. The LOI is annotated for every such sub-speaker turn. In order to get an impression of a subject’s character and behavior prior to the actual annotation, the annotators had to watch approximately five minutes of a subject’s video. As the focus of interest based annotation lies on the sub-speaker turn, each of those had to be viewed at least once to find out the LOI displayed by the subject. Five levels of interest were distinguished:

- LOI−2 Disinterest (subject is tired of listening and talking about the topic, is totally passive, and does not follow)
- LOI−1 Indifference (subject is passive, does not give much feedback to the experimenter’s explanations, and asks unmotivated questions, if any)
- LOI0 Neutrality (subject follows and participates in the discourse; it cannot be recognized if she/he is interested or indifferent in the topic)
- LOI+1 Interest (subject wants to discuss the topic, closely follows the explanations, and asks questions)
- LOI+2 Curiosity (strong wish of the subject to talk and learn more about the topic).

To avoid different interpretations of the LOI names, the annotators used the LOI values from −2 to 2 rather than the terms (such as *Neutrality* or *Curiosity*) when assigning their labels. Otherwise inconsistencies could have potentially occurred, since terms like *Indifference* and *Neutrality* might be interpreted differently or even used synonymously. The inter-labeler agreement can be seen as sufficiently high (κ-value of 0.66, see [20]).

Further, the spoken content has been transcribed, and long pause, short pause, and non-linguistic vocalizations have been labeled. These vocalizations comprise breathing (452), consent (325), hesitation (1,147), laughter (261),

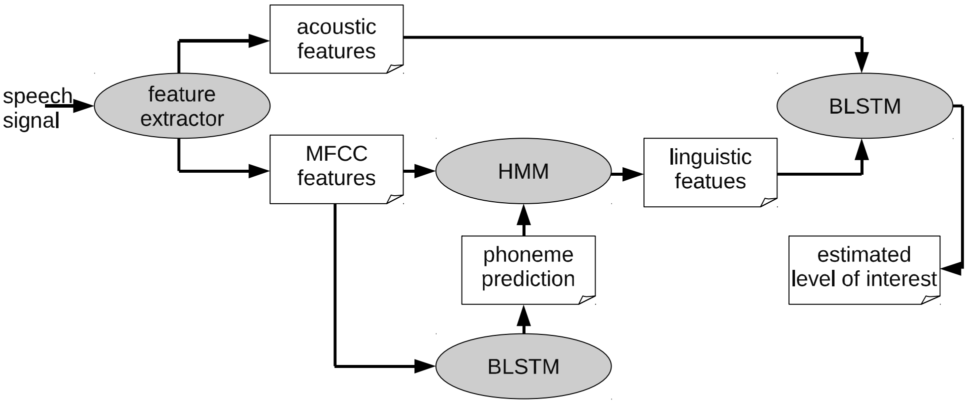

Fig. 1 System architecture for acoustic-linguistic interest recognition

and coughing, other human noise (716). There is a total of 18,581 spoken words, and 23,084 word-like units including 2,901 non-linguistic vocalizations (19.5%). In summary, the overall annotation contains the spoken content, non-linguistic vocalizations, individual annotator tracks, and mean LOI (per sub-speaker-turn segment).

For the Interspeech 2010 Paralinguistic Challenge, the ground truth is established by shifting to a continuous scale obtained by averaging the single annotator LOI. In accordance with the scaling applied in other corpora (e.g., [6]), the original LOI scale reaching from LOI-2 to LOI$+2$ is mapped to the interval from -1 to 1. Note that the level of interest introduced herein is highly correlated to arousal. However, at the same time there is an obvious strong correlation to valence, as e.g., boredom has a negative valence, while strong interest is characterized by positive valence. The annotators however labeled interest in the common sense, thus comprising both aspects.

The speech data from the 21 speakers (3,880 sub-speaker-turns) were split into a training, development, and test set. Splitting was conducted in a speaker independent way trying to achieve the best possible balance with respect to gender, age, and ethnicity. The training set consists of 1,512 sub-speaker-turns and 51.7 minutes of speech, respectively, and comprises four female and four male speakers, while the development set contains 1,161 sub-speaker-turns, corresponding to 43.1 minutes of speech (three female and three male speakers). The test set includes 1,207 sub-speaker-turns and 42.7 minutes of speech, respectively (three female and four male speakers).

3 Long Short-Term Memory

This section outlines the principle of the Long Short-Term Memory RNNs that are used for context-sensitive interest recognition in Sect. 5 as well as for phoneme prediction in Sect. 4.2. The architecture of the whole acoustic-linguistic interest recognition system is shown in Fig. 1: A feature extractor provides MFCC features to a BLSTM network which computes a phoneme prediction. Together with the MFCC features, those phoneme predictions are decoded by a multi-stream HMM which outputs linguistic features. Both, linguistic features and acoustic features are processed by a second BLSTM network which infers the final level of interest prediction.

The automatic prediction of a user's level of interest as investigated in this article profits from classification architectures that can access and model long-range context since the level of interest is expected to evolve slowly over time, with past observations potentially influencing the current prediction. The *number* of past (and possibly future) speech turns which should be used to obtain enough context for reliably estimating the level of interest without affecting the capability of also detecting sudden changes of the speaker's affective state is hard to determine [21, 24]. Thus, a classifier that is able to *learn* the amount of context is a promising alternative to manually defining fixed time windows for interest recognition. Static techniques such as Support Vector Machines do not explicitly model context between turns but rely, e.g., on aggregating observations using Multi-Instance Learning techniques [22]. Other classifiers such as neural networks are able to model a certain amount of context by using cyclic connections. These so-called recurrent neural networks (RNN) 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 resulted in 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 [9]). One of the most effective solutions to this problem is the Long Short-Term Memory (LSTM) architecture [10], which is able to store information in linear memory cells over a longer period of time. LSTM networks can learn the optimal amount of contextual information relevant for the classification task and thus are well-suited for context-sensitive interest recognition.

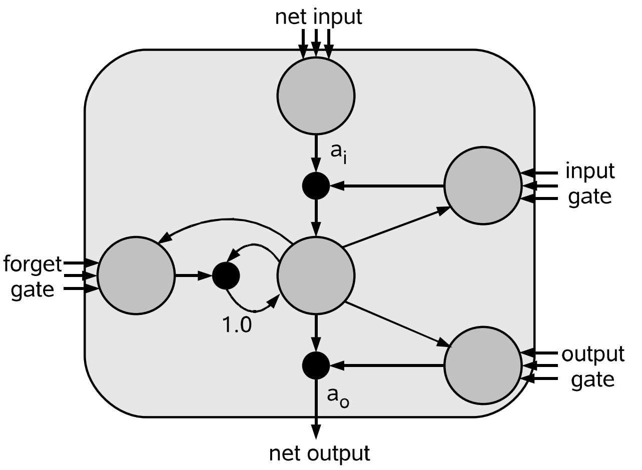

Fig. 2 LSTM memory block consisting of one memory cell: the input, output, and forget gates collect activations from inside and outside the block which control the cell through multiplicative units (depicted as *small circles*); input, output, and forget gate scale input, output, and internal state respectively; a_i and a_o denote activation functions; the recurrent connection of fixed weight 1.0 maintains the internal state

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. 2). The overall effect is to allow the network to store and retrieve information over long periods of 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 [25], 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.

Combining bidirectional networks with LSTM gives bidirectional LSTM [5], which has demonstrated excellent performance in various applications like handwriting recognition [4], music information retrieval [28], speech recognition [33], keyword spotting [31], and emotion recognition [32, 35, 36].

Note that bidirectional processing is rather suited for offline information retrieval than for fully incremental real-time systems since it requires future information. Nevertheless exploiting future context can also be an interesting aspect for on-line systems that aim at refining past predictions once more (bidirectional) context is available. Often a short look-ahead buffer is enough to profit from a limited amount of 'future' information, as for example when applying triphone models for speech recognition, modeling the coarticulation that occurs when preceding and successive phonemes affect the pronuncation of the current phoneme.

4 Feature Extraction

In the following sections we describe the features that we apply for classification. We will review the set of acoustic features that has been proposed in [23] in order to define a unified feature set that can be used for comparing the accuracy of different classification approaches (Sect. 4.1). Unlike the baseline interest predictor that had been introduced in [23] and is exclusively based on acoustic descriptors, the regression technique applied in this article also makes use of linguistic features and thus requires an ASR module recognizing spoken content and non-linguistic vocalizations such as *laughing*. In Sect. 4.2 we outline our multi-stream speech recognizer providing linguistic features. Finally, in Sect. 4.3 we describe the feature selection technique which we apply to reduce the dimensionality of the feature space.

4.1 Acoustic Features

The acoustic features applied in Sect. 5 correspond to the baseline feature set of the Interspeech 2010 Paralinguistic Challenge [23]. They are extracted via our real-time speech analysis toolbox openSMILE [2] which has emerged as a widely-adopted audio feature extractor used in various studies on affective computing and paralinguistic information extraction [1, 8, 15, 17].

1,582 acoustic features are obtained in total by systematic 'brute-force' feature generation in three steps: first, the 38 low-level descriptors (LLD) shown in Table 1 (left column) are extracted at 100 frames per second with varying window type and size (Hamming and 25 ms, respectively, for all but pitch which is extracted using a Gaussian window and a window size of 60 ms) and smoothed by simple moving average low-pass filtering with a window length of three frames. Next, their first order regression coefficients are added. Then, 21 functionals are applied (see Table 1, right column) to each low-level feature stream in order to capture time-varying information in a fixed-length static feature vector for each instance in the database. Note that 16 zero-information features (e.g., minimum $F0$, which is always zero) are discarded. Finally, the two single features 'number of pitched segments' and turn duration are added.

Table 1 The official 1,582-dimensional acoustic feature set of the Interspeech 2010 Paralinguistic Challenge: 38 low-level descriptors with regression coefficients, 21 functionals. Abbreviations: DDP: difference of difference of periods, LSP: line spectral pairs, Q/A: quadratic, absolute

Descriptors	Functionals
PCM loudness	Max./min. (position)
MFCC [0–14]	Arith. mean, std. deviation
log Mel Freq. Band [0–7]	Skewness, kurtosis
LSP Frequency [0–7]	Lin. regression coeff. 1/2
F0 by Sub-Harmonic Sum.	Lin. regression error Q/A
F0 Envelope	Quartile 1/2/3
Voicing Probability	Quartile range 2–1, 3–2, 3–1
Jitter local	Percentile 1/99
Jitter DDP	Percentile range 99–1
Shimmer local	Up-level time 75/90

4.2 Linguistic Features

This section briefly outlines the multi-stream BLSTM-HMM ASR system we use to generate linguistic features. In general, spontaneous, disfluent speech can lead to extremely high error rates when applying conventional HMM-based speech recognizers. Thus, even if an ASR system just serves as linguistic feature extractor, challenging scenarios such as the TUM AVIC interactions require more advanced strategies in order to obtain the best possible ASR performance and linguistic features, respectively. For our experiments we applied the technique presented in [34]. It was shown that this multi-stream model architecture is particularly suited for robust speech recognition in challenging scenarios (conversational speech, emotional coloring of speech, background noise, etc.). The main idea of this technique is to enable improved recognition accuracies by incorporating context-sensitive phoneme predictions generated by a Bidirectional Long Short-Term Memory network (see Sect. 3) into the speech decoding process.

The structure of our multi-stream decoder can be seen in Fig. 3: s_t and x_t represent the HMM state and the acoustic (MFCC) feature vector, respectively, while b_t corresponds to the discrete phoneme prediction of the BLSTM network (shaded nodes). Squares denote observed nodes and white circles represent hidden nodes. In every time frame t the HMM uses two independent observations: the MFCC features x_t and the BLSTM phoneme prediction feature b_t. The vector x_t also serves as input for the BLSTM, whereas the size of the BLSTM input layer i_t corresponds to the dimensionality of the acoustic feature vector. The vector o_t contains one probability score for each of the P different phonemes at each time step and b_t is the index of the most likely phoneme. In every time step the BLSTM generates a phoneme prediction and the HMM models $x_{1:T}$ and $b_{1:T}$ as two independent data streams.

The applied real-time LSTM-based phoneme predictor is publicly available as part of our on-line speech feature extraction engine openSMILE [2].

Via early fusion, we fuse linguistic information extracted by the BLSTM-HMM speech recognizer with the supra-segmental acoustic features described in Sect. 4.1. To obtain linguistic feature vectors from the ASR output, a standard Bag-of-Words (BoW) technique is employed: for each word in a segment, the term frequency (TF) is computed (see [12]). Only words with a minimum term frequency of two throughout the training set are considered (152 words). A vector space representation of the word string is built from the word's TF values.

4.3 Feature Selection

In order to reduce the size of the resulting (acoustic-linguistic) feature space, we conduct a cyclic Correlation based Feature Subset Selection (CFS) based on the TUM AVIC training set. 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.

For correlation measurement, the symmetrical uncertainty coefficient is used (as described in [7]). To avoid an exhaustive search in the feature space a greedy hill climbing forward search is applied [29]. In this heuristic search algorithm, each feature is tentatively added to the feature subset. 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. As termination criterion we considered a maximum of five non-improving nodes before terminating the greedy hill climbing forward search.

Table 2 gives an overview over the acoustic features selected by CFS on the TUM AVIC training set. We indicate the number (#) of features that were selected, and their percentage with respect to the full 1,582-dimensional Paralinguistic Challenge feature set, which is reduced to 92 features (about 6%).

First, in Table 2a, we distinguish the features by their types. Judging from the percentages, it can be seen that a broad range of prosodic, spectral, and voice quality features are correlated with the mean level of interest, with no considerable difference between these three groups. Overall, it is striking that regression coefficients (Δ LLD) seem to contribute equally to information about the mean level of interest as the LLDs themselves.

Second, we investigate the contribution of different types of functionals in Table 2b, concluding that especially the percentiles seem to carry valuable information: Note that the

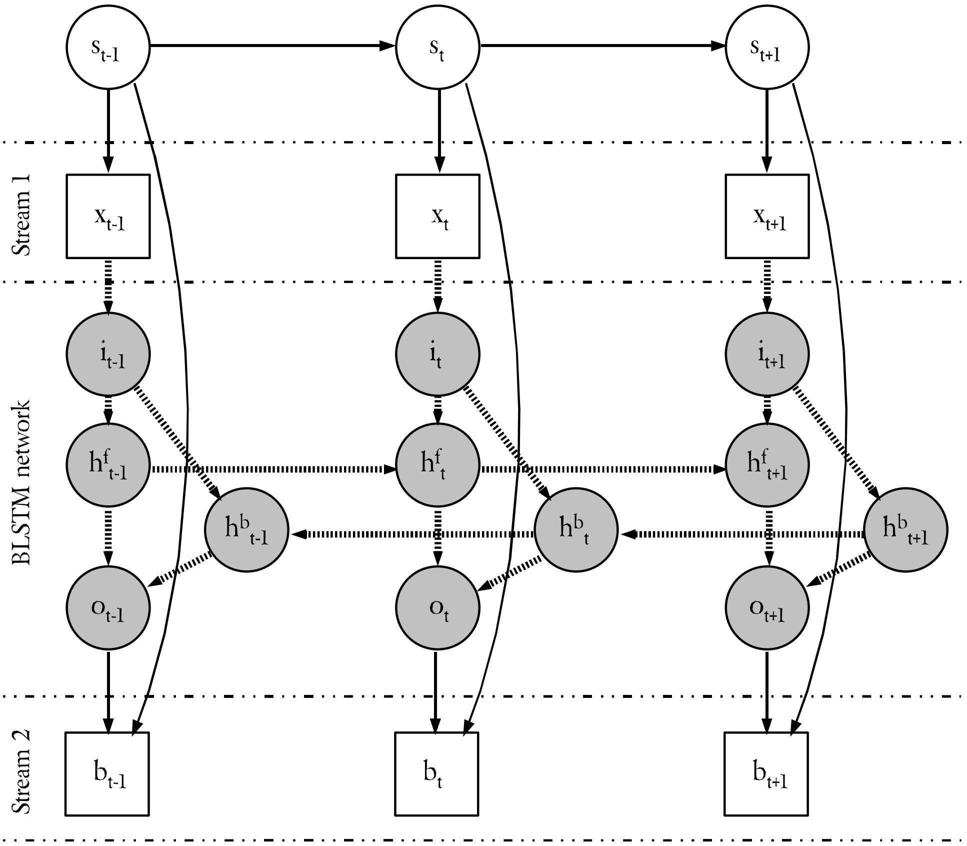

Fig. 3 Architecture of the multi-stream BLSTM-HMM decoder: s_t: HMM state, x_t: acoustic feature vector, b_t: BLSTM phoneme prediction feature, i_t, o_t, h_t^f/h_t^b: input, output, and hidden nodes of the BLSTM network.

Table 2 Acoustic feature selection (CFS) on the TUM AVIC training set: number and percentage of features selected from the 1,582-dimensional Paralinguistic Challenge feature set, (a) by type of LLD (prosodic, spectral, voice quality), and (b) by type of functional

[#]	LLD	Δ LLD	Total
(a) By type of LLD			
Prosodic	5/63 (8%)	0/61 (0%)	5/124 (4%)
Spectral	37/651 (6%)	39/651 (6%)	76/1302 (6%)
Voice q.	6/78 (8%)	5/78 (6%)	11/156 (7%)
Total	48/792 (6%)	44/790 (6%)	92/1582 (6%)
(b) By type of functionals			
Extremes	10/182 (5%)	10/182 (5%)	20/364 (5%)
Regression	7/152 (5%)	9/152 (6%)	16/304 (5%)
Moments	10/152 (7%)	9/152 (6%)	19/304 (6%)
Percentiles	16/228 (7%)	15/228 (7%)	31/456 (7%)
Duration	5/76 (7%)	1/76 (1%)	6/152 (4%)
Total	48/792 (6%)	44/790 (6%)	92/1582 (6%)

Table 3 Results of acoustic and linguistic feature selection (CFS) on the TUM AVIC training set. The number and percentage of features selected from the union of the 1,582-dimensional acoustic feature set and the BoW feature vector (152 features) are shown by type of LLD

# Selected LLD + Δ LLD	
Prosodic	5/124 (4%)
Spectral	66/1302 (5%)
Voice quality	12/156 (8%)
Σ acoustic	83/1582 (5%)
BOW	38/147 (26%)
BOW non-ling.	2/5 (40%)
Σ linguistic	40/152 (26%)
Total	123/1734 (7%)

Extremes category contains the 1- and 99-percentile instead of actual minimum and maximum for robustness against outliers.

Next, in Table 3, we summarize the results of CFS-based feature selection from the union of the acoustic and the linguistic (BoW) feature set. For the sake of clarity, we do not consider regression coefficients separately. Regarding the selection of acoustic features from the joint feature set as opposed to the acoustic set only (Table 2), figures are comparable, yet slightly less features are selected in total (83 vs. 92). Interestingly, and in contrast to acoustic features, a large share of the original BoW feature space is kept (40 of 152, or 26%), which particularly includes two of the five features corresponding to the non-linguistic vocalizations *consent* ("mhm") and *laughter*. Other selected BoW features correspond to set phrases such as *oh*, *yeah*, or *good*, but also to words that can be judged as being relevant for the corpus-specific car presentation scenario such as *hybrid*, *buy*, or *gas*.

5 Experiments and Results

5.1 ASR Configuration and Training

As outlined in Sect. 4.2, our ASR module for linguistic feature extraction combines the HMM-based decoding with context-sensitive BLSTM-based phoneme prediction. We trained the multi-stream ASR system on utterances from the training and the development partition of the TUM AVIC corpus. As ASR features x_t we used MFCCs 1 to 12 including logarithmic energy together with first and second order regression coefficients. To compensate for stationary noise effects, we applied cepstral mean normalization. Since the BLSTM network for phoneme prediction was trained on framewise phoneme targets, we used an HMM system to obtain phoneme borders via forced alignment. The network consisted of three hidden layers (per input direction) with a size of 78, 128, and 94 hidden units, respectively. Thereby each memory block contained one memory cell. For training of the BLSTM-based phoneme predictor we used a learning rate of 10^{-5} and a momentum of 0.9. As a common means to improve generalization for RNNs, we added zero mean Gaussian noise with standard deviation 0.6 to the inputs during training. 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. We trained the networks on the standard (CMU) set of 39 different English phonemes and included targets for *silence*, *short pause*, and *garbage* as well as for the non-linguistic vocalizations *breathing*, *coughing*, *hesitation*, *laughing*, and *consent* (like "mhm").

Each phoneme of the underlying HMM system is represented by three states (left-to-right HMMs) with 16 Gaussian mixtures. HMMs corresponding to non-linguistic vocalizations consisted of nine states. The initial monophone models consisted of one Gaussian mixture per state and were trained using four iterations of embedded Baum-Welch re-estimation. After that, the monophones were mapped to tied-state cross-word triphone models with shared state transition probabilities. Two Baum-Welch iterations were performed for re-estimation of the triphone models. Finally, the number of mixture components of the triphone models was increased to 16 in four successive rounds of mixture doubling and re-estimation. Both, the multi-stream acoustic models and a back-off bigram language model were trained on the TUM AVIC training and development set (vocabulary size of 1.9 k). Applying the multi-stream model, we obtain a word accuracy of 28.9% (compared to 25.8% when using only the MFCC feature stream), which is in the range of typical accuracies for such challenging ASR scenarios involving accented and spontaneous speech with a high number of out-of-vocabulary events [11].

5.2 Neural Network Architectures for Interest Recognition

We evaluated four different neural network architectures with respect to their suitability for speech-based interest recognition: conventional recurrent neural networks, bidirectional recurrent neural networks (BRNN), LSTM networks, and Bidirectional LSTM networks. The number of input nodes corresponds to the number of selected acoustic or combined acoustic-linguistic features. BLSTM networks were composed of 32 memory blocks per input direction, while LSTM networks consisted of 64 memory blocks. Similarly to the BLSTM phoneme predictor, all memory blocks of the interest prediction (B)LSTMs were composed of one memory cell. RNN classifiers had a hidden layer of size 32 while BRNNs consisted of 16 hidden nodes per input direction. All networks had one (regression) output node whose activation represents the predicted level of interest.

As for the phoneme predictor, we improved generalization by adding Gaussian noise to the inputs during training (standard deviation of 1.2). Note that all input features were z-normalized before being processed by the networks. Means and standard deviations for z-normalization were computed from the training set. The remaining configurations are similar to the parameterization for the phoneme predictor, however, for interest recognition we applied resilient propagation (rProp) [18] instead of standard backpropagation through time. Learning rate and momentum were set to 10^{-5} and 0.9, respectively.

5.3 Results

Table 4 shows the results obtained on the Interspeech 2010 Paralinguistic Challenge (more precisely the *Affect Sub-Challenge*) when applying the different context-sensitive neural network architectures. In conformance with [23], we chose the cross correlation (CC) between the ground truth level of interest and the predicted level of interest as evaluation criterion. We do *not* report the mean linear error (MLE), since the MLE strongly depends on the variance of the ground truth labels and is hardly suited for revealing the accuracy of the predictions. As an example, when evaluating a ('dummy') classifier that always predicts the mean of the training set ground truth labels, we obtain an MLE of 0.148 (which is only 0.002 below the MLE reported in [23]) while we get a CC of zero.

All results reflect the recognition performance on the TUM AVIC test set, when training the predictors on the

Table 4 Results for interest recognition as defined in the Affect Sub--Challenge [23]: cross correlation obtained for different network architectures when using either acoustic (Ac.) or combined acoustic-linguistic (Ac. + Ling.) information; baseline results reported in [23] when applying unpruned REP-Trees with and without correlation-based feature selection (CFS)

Classifier	CFS	Cross correlation	
		Ac.	Ac. + Ling.
BLSTM	Yes	0.442	0.475
LSTM	Yes	0.431	0.459
BRNN	Yes	0.406	0.438
RNN	Yes	0.422	0.439
REP-Trees	Yes	0.439	0.435
REP-Trees [23]	No	0.421	0.423

training and development partition of the TUM AVIC corpus. Using only the training set did not lead to satisfying results since our neural network architectures require a comparatively large amount of training data for generalization. Incorporating linguistic information leads to higher cross correlations for all network architectures which is in line with previous studies on speech based affect recognition (e.g., [36]). The best performance can be obtained when applying Bidirectional Long Short-Term Memory networks processing both, acoustic and linguistic features (CC of 0.475). Bidirectional LSTM modeling gives slightly better results than unidirectional LSTM which indicates that also future information (if available) can be efficiently exploited for interest recognition. The performance difference between LSTM-based architectures and conventional RNN techniques reveals that the ability to model long-term temporal context is beneficial for our classification task.

For comparison, also the Paralinguistic Challenge baseline result (CC of 0.421, obtained with unpruned REP-Trees in Random-Sub-Space meta-learning [23]) is shown in Table 4. The REP-Trees approach profits from feature selection via CFS but cannot compete with the BLSTM technique. Our results are even significantly better than the highest cross correlation that has ever been reported for the Affect Sub-Challenge so far (CC of 0.428 using acoustic and linguistic information [11]).

We believe that a large part of the performance gain with respect to previous studies on interest recognition can be attributed to our strategy of not looking at single utterances in isolation but modeling how the user's interest evolves over time. This is in conformance with the way humans judge the affective state of others: They observe people over longer time spans and tend to place utterances (or, more general, behaviors) into context. Another reason for the good performance of our approach is the applied feature selection which leads to a reduction of the feature space and of the number of network input nodes, respectively. Smaller networks processing a moderate number of inputs are generally easier to train with a limited amount of training material.

6 Conclusion and Outlook

This article presented a fully automatic approach towards recognition of a user's level of interest, based on information extracted from the speech signal. We showed how a context-sensitive neural network architecture based on the Long Short-Term Memory principle can be applied for improved assessment of interest exploiting both, acoustic (i.e., prosodic, spectral, and voice quality) and linguistic information (i.e., the spoken content including non-linguistic vocalizations). Unlike previous studies extracting linguistic features from the ground truth transcription of the spoken content, we focus on processing the ASR output, which of course is error-prone but reflects the conditions a real-life interest recognition system has to face. Since recognition of spontaneous, conversational, and potentially emotionally colored speech is a challenge in itself, we apply an advanced multi-stream ASR system that exploits the LSTM principle for enhanced phoneme recognition and integrates inferred phoneme predictions into an HMM framework. Using this acoustic-linguistic interest recognition technique, we were able to outperform the baseline recognizer as well as all other classification techniques that had been proposed for the Interspeech 2010 Paralinguistic Challenge—an initiative to make recognition results of different research teams comparable by defining unified test conditions.

The system proposed in this article can be seen as a versatile speech-based interest recognizer that does not have to be adapted to a certain speaker or to the characteristics of a certain speaker. This, however, implies that our system cannot capture that different users might express interest in different ways. User-profiled interest estimation in turn would mean that user-specific models have to be used. Yet, in real-life applications the availability of user profiles is an unrealistic assumption and general (or 'average') models that can cope with unknown users are of higher importance. Another limitation is that the user has to express interest or disinterest verbally, i.e., an interested user who is listening without giving verbal feedback will of course not be captured by a speech-based system.

Our results indicate that the prediction of human interest can best be performed when applying model architectures that consider contextual information instead of classifying a speech turn in isolation. Similar findings have been published for other affective dimensions, such as valence and arousal (see [35, 36]). This raises the question whether overall recognition accuracies can be improved when *jointly* predicting multiple emotional dimensions—including the level of interest—via multi-task learning. Another interesting direction for future research is to examine framewise multimodal prediction in combination with hybrid fusion techniques [30] which are able to model asynchronities between multiple modalities or input streams such as acoustic, linguistic, or video features. Also alternative ways for determining the ground truth interest level of an utterance are thinkable for future experiments. For the TUM AVIC example, the 'true' level of interest could be determined by self-report methods or by asking the participants to summarize the content of the presentation and derive the level of interest from the quality of the summary which in turn indicates how attentive or interested the user has been.

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

References

1. Devillers L, Vaudable C, Chastagnol C (2010) Real-life emotion-related states detection in call centers: a cross-corpora study. In: Proc of interspeech, Makuhari, Japan, pp 2350–2353
2. Eyben F, Wöllmer M, Schuller B (2010) openSMILE—the Munich versatile and fast open-source audio feature extractor. In: Proc of ACM multimedia, Firenze, Italy, pp 1459–1462
3. Gatica-Perez D, McCowan I, Zhang D, Bengio S (2005) Detecting group interest-level in meetings. In: Proc of ICASSP, Philadelphia, USA, pp 489–492
4. Graves A, Fernandez S, Liwicki M, Bunke H, Schmidhuber J (2008) Unconstrained online handwriting recognition with recurrent neural networks. Adv Neural Inf Process Syst 20:1–8
5. Graves A, Fernandez S, Schmidhuber J (2005) Bidirectional LSTM networks for improved phoneme classification and recognition. In: Proc of ICANN, Warsaw, Poland, pp 602–610
6. Grimm M, Kroschel K, Narayanan S (2008) The Vera am Mittag german audio-visual emotional speech database. In: Proc of ICME, Hannover, Germany, pp 865–868
7. Hall MA (1999) Correlation-based feature selection for machine learning. PhD thesis, University of Waikato
8. Hassan A, Damper RI (2010) Multi-class and hierarchical SVMs for emotion recognition. In: Proc of interspeech, Makuhari, Japan, pp 2354–2357
9. Hochreiter S, Bengio Y, Frasconi P, Schmidhuber J (2001) Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In: Kremer SC, Kolen JF (eds) A field guide to dynamical recurrent neural networks. IEEE Press, New York, pp 1–15
10. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
11. Jeon JH, Xia R, Liu Y (2010) Level of interest sensing in spoken dialog using multi-level fusion of acoustic and lexical evidence. In: Proc of interspeech, Makuhari, Japan, pp 2802–2805
12. Joachims T (1998) Text categorization with support vector machines: learning with many relevant features. In: Proc of ECML, Chemnitz, Germany, pp 137–142
13. Kennedy L, Ellis D (2003) Pitch-based emphasis detection for characterization of meeting recordings. In: Proc of ASRU, Virgin Islands, pp 243–248
14. Mota S, Picard R (2003) Automated posture analysis for detecting learner's interest level. In: Proc of workshop on CVPR for HCI, Madison, pp 49–55
15. Nguyen P, Tran D, Huang X, Sharma D (2010) Automatic classification of speaker characteristics. In: Proc of ICCE, pp 147–152
16. Pentland A, Madan A (2005) Perception of social interest. In: Proc of IEEE international conference on computer vision, workshop on modeling people and human interaction (ICCV-PHI), Beijing, China
17. Pfister T, Robinson P (2010) Speech emotion classification and public speaking skill assessment. In: Salah A, Gevers T, Sebe N, Vinciarelli A (eds) Human behavior understanding. Lecture notes in computer science, vol 6219. Springer, Berlin, pp 151–162
18. Riedmiller M, Braun H (1993) A direct adaptive method for faster backpropagation learning: the RPROP algorithm. In: Proc of IEEE international conference on neural networks, pp 586–591
19. Schröder M, Cowie R, Heylen D, Pantic M, Pelachaud C, Schuller B (2008) Towards responsive sensitive artificial listeners. In: Proc of 4th intern workshop on human-computer conversation, Bellagio, Italy, pp 1–6
20. Schuller B, Müller R, Eyben F, Gast J, Hörnler B, Wöllmer M, Rigoll G, Höthker A, Konosu H (2009) Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. Image Vis Comput J 27(12):1760–1774. Special issue on visual and multimodal analysis of human spontaneous behavior
21. Schuller B, Rigoll G (2006) Timing levels in segment-based speech emotion recognition. In: Proc of interspeech, Pittsburgh, USA, pp 1818–1821
22. Schuller B, Rigoll G (2009) Recognising interest in conversational speech—comparing bag of frames and supra-segmental features. In: Proc of interspeech, Brighton, UK, pp 1999–2002
23. Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Müller C, Narayanan S (2010) The interspeech 2010 paralinguistic challenge. In: Proc of interspeech, Makuhari, Japan, pp 2794–2797
24. Schuller B, Vlasenko B, Minguez R, Rigoll G, Wendemuth A (2007) Comparing one and two-stage acoustic modeling in the recognition of emotion in speech. In: Proc of ASRU, Kyoto, Japan, pp 596–600
25. Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45:2673–2681
26. Shriberg E (2005) Spontaneous speech: how peoply really talk and why engineers should care. In: Proc of interspeech, Lisbon, Portugal, pp 1781–1784
27. Stiefelhagen R, Yang J, Waibel A (2002) Modeling focus of attention for meeting indexing based on multiple cues. IEEE Trans Neural Netw 13(4):928–938
28. Weninger F, Durrieu JL, Eyben F, Richard G, Schuller B (2011) Combining monoaural source separation with long short-term memory for increased robustness in vocalist gender recognition. In: Proc of ICASSP, Prague, Czech Republic
29. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco
30. Wöllmer M, Al-Hames M, Eyben F, Schuller B, Rigoll G (2009) A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams. Neurocomputing 73(1-3):366–380
31. Wöllmer M, Eyben F, Graves A, Schuller B, Rigoll G (2010) Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework. Cognit Comput 2(3):180–190
32. Wöllmer M, Eyben F, Reiter S, Schuller B, Cox C, Douglas-Cowie E, Cowie R (2008) Abandoning emotion classes—towards continuous emotion recognition with modelling of long-range dependencies. In: Proc of interspeech, Brisbane, Australia, pp 597–600

33. Wöllmer M, Eyben F, Schuller B, Rigoll G (2010) Recognition of spontaneous conversational speech using long short-term memory phoneme predictions. In: Proc of interspeech, Makuhari, Japan, pp 1946–1949
34. Wöllmer M, Eyben F, Schuller B, Rigoll G (2011) A multi-stream ASR framework for BLSTM modeling of conversational speech. In: Proc of ICASSP, Prague, Czech Republic
35. Wöllmer M, Metallinou A, Eyben F, Schuller B, Narayanan S (2010) Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In: Proc of interspeech, Makuhari, Japan, pp 2362–2365
36. Wöllmer M, Schuller B, Eyben F, Rigoll G (2010) Combining long short-term memory and dynamic Bayesian networks for incremental emotion-sensitive artificial listening. IEEE J Sel Top Signal Process 4(5):867–881
37. Zeng Z, Pantic M, Rosiman GI, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58