Acoustic-linguistic recognition of interest in speech with bottleneck-BLSTM nets

Martin Wöllmer, Felix Weninger, Florian Eyben, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Wöllmer, Martin, Felix Weninger, Florian Eyben, and Björn Schuller. 2011. "Acoustic-linguistic recognition of interest in speech with bottleneck-BLSTM nets." In *INTERSPEECH 2011 - 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, edited by Piero Cosi, Renato De Mori, Giuseppe Di Fabbrizio, and Roberto Pieraccini, 77–80. ISCA Archive. https://doi.org/10.21437/Interspeech.2011-20.

Acoustic-Linguistic Recognition of Interest in Speech with Bottleneck-BLSTM Nets

Martin Wöllmer, Felix Weninger, Florian Eyben, Björn Schuller

Institute for Human-Machine Communication, Technische Universität München, Germany

[woellmer,weninger,eyben,schuller]@tum.de

10.21437/Interspeech.2011-20

Abstract

This paper proposes a novel technique for speech-based interest recognition in natural conversations. We introduce a fully automatic system that exploits the principle of bidirectional Long Short-Term Memory (BLSTM) as well as the structure of so-called bottleneck networks. BLSTM nets are able to model a self-learned amount of context information, which was shown to be beneficial for affect recognition applications, while bottleneck networks allow for efficient feature compression within neural networks. In addition to acoustic features, our technique considers linguistic information obtained from a multi-stream BLSTM-HMM speech recognizer. Evaluations on the TUM AVIC corpus reveal that the bottleneck-BLSTM method prevails over all approaches that have been proposed for the Interspeech 2010 Paralinguistic Challenge task.

Index Terms: affective computing, interest recognition, recurrent neural networks

1. Introduction

Detecting whether a user is interested or disinterested can be relevant for many applications of Human-Computer Interaction, including sales and advertisement systems, virtual guides, or conversational agents. Recently investigated use-cases for automatic interest recognition comprise topic switching in infotainment or customer service systems [1], meeting analysis, and tutoring systems [2]. In the light of this growing amount of research on interest-related affective computing, the organizers of the Interspeech 2010 Paralinguistic Challenge [3] 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. Participants used the Audiovisual Interest Corpus recorded at the Technische Universität München ("TUM AVIC") [1]. It contains 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 was to predict a speaker's level of interest by suited regression techniques.

In this paper, we attempt to exploit contextual information for enhanced acoustic-linguistic interest recognition by employing a context-sensitive neural network architecture. Building on our recent studies on the incorporation of long-range context knowledge via Long Short-Term Memory (LSTM) recurrent neural networks [4] for emotion recognition applications [5, 6], we apply bidirectional LSTM (BLSTM) networks to model how the user's interest level evolves over time. LSTM networks overcome the so-called *vanishing gradient problem* [7] which makes it difficult to learn long-range context via conventional recurrent neural networks (RNN). In contrast to previous LSTM-based emotion recognition systems which contain one hidden layer [6], we design *bottleneck*-BLSTM networks by using three hidden layers with a narrow middle layer (the so-called 'bottleneck'). Bottleneck networks have recently been introduced for automatic speech recognition (ASR) where they can be applied for feature dimensionality reduction within Tandem systems [8], i. e., speech recognizers that use RNNs or multi-layer perceptrons (MLP) to generate features that are modeled with Hidden Markov Models (HMM). For our interest recognition system, we combine the bottleneck principle with the BLSTM technique and generate a compact feature representation within the BLSTM network. In addition to acoustic features, the bottleneck-BLSTM network processes linguistic information obtained from an ASR module. Since ASR in conversational speech scenarios tends to be more challenging than, e. g., the recognition of read speech, we apply a multi-stream BLSTM-HMM speech recognizer which has shown good performance in challenging spontaneous speech scenarios [9]. 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 structure of this paper is as follows: Section 2 introduces the TUM AVIC corpus, Section 3 explains the principle of bottleneck-BLSTM modeling, Section 4 contains details about the feature extraction, and Section 5 shows experimental results. Conclusions are drawn in Section 6.

2. TUM AVIC Corpus

Our experiments are based on the TUM AVIC corpus [1] which has also been used for the Affect Sub-Challenge of the Interspeech 2010 Paralinguistic Challenge [3]. 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.

The 'level of interest' (LOI) is annotated for every turn using five levels of interest from disinterest to curiosity (LOI -2, -1, 0, 1, 2). Further, the spoken content as well as non-linguistic vocalizations have been transcribed. For the Interspeech 2010 Paralinguistic Challenge, the ground truth has been established by shifting to a continuous scale obtained by averaging the single annotator LOI. In accordance with the scaling applied in other corpora, the original LOI scale reaching from -2 to +2 is

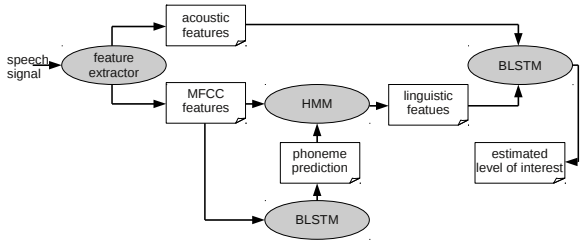

Figure 1: *System architecture for acoustic-linguistic interest recognition.*

mapped to the interval from -1 to 1.

The speech data from the 21 speakers (3 880 turns) were split into speaker independent training, development, and test sets. The training set consists of 1 512 turns and 51.7 minutes of speech, respectively, and comprises four female and four male speakers, while the development set contains 1 161 turns, corresponding to 43.1 minutes of speech (three female and three male speakers). The test set includes 1 207 turns and 42.7 minutes of speech, respectively (three female and four male speakers). More details on the TUM AVIC corpus can be found in [3].

3. Bottleneck-BLSTM Nets

Building on recent successes of LSTM-based affective computing and speech recognition [5, 6, 9], we apply Long Short-Term Memory RNNs for context-sensitive interest recognition (Section 5) as well as for phoneme prediction (see Section 4.2). The architecture of the whole acoustic-linguistic interest recognition system is shown in Figure 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 paper 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. 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.

Recurrent neural networks are able to model a certain amount of context by using cyclic connections. 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 [7]). An elegant solution to this problem is the Long Short-Term Memory (LSTM) architecture [4], which is able to store information in linear memory cells over a longer period of time. LSTM nets can learn the optimal amount of contextual information relevant for the classification task and thus are well-suited for context-sensitive

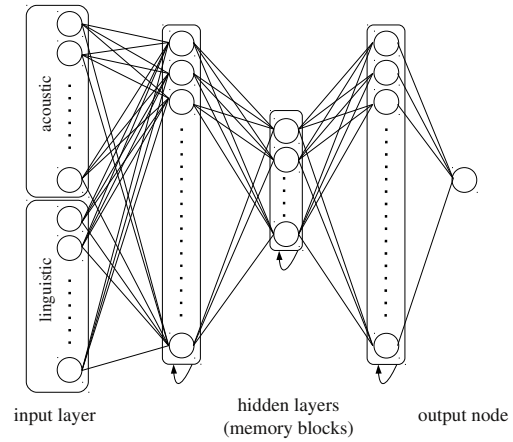

Figure 2: *Structure of the bottleneck networks used for interest recognition.*

interest recognition.

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 (input, output, and forget gates). These gates perform functions analogous to read, write, and reset operations. The overall effect is to allow the network to store and retrieve information over long periods of time. For more details on LSTM networks see [10], for example.

Bidirectional networks presume that both, past and future context information can be used. A bidirectional RNN (or LSTM) consists of two separate recurrent hidden layers which 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. Note that bidirectional processing is rather suited for off-line 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.

This paper presents a first attempt to combine the principle of LSTM with so-called bottleneck network architectures. Bottleneck MLPs or RNNs consist of (at least) three hidden layers with a narrow layer in the middle (the bottleneck). Those networks have recently been introduced for feature generation in Tandem speech recognition systems [8]. Instead of using phoneme posterior features corresponding to the output activations of an MLP trained on phoneme targets, bottleneck ASR systems apply features that are obtained from the linear outputs of the neurons in the bottleneck layer, i. e., only the first two hidden layers are involved during feature extraction. This offers the advantage that by choosing the size of the bottleneck layer, the dimensionality of the feature vector can be defined. Thus, the network implicitly performs dimensionality reduction and generates decorrelated and compressed features – independent of the number of training targets and without the need for explicit decorrelation and dimensionality reduction techniques such as Principal Component Analysis (PCA). Unlike static techniques based on PCA (or MLPs), combining LSTM and bottleneck architectures enables *context-sensitive* feature compression.

For interest recognition, we apply five-layer bottleneck-

LSTMs as shown in Figure 2. The networks are composed of an input layer whose size corresponds to the dimensionality of the acoustic-linguistic feature vector, three hidden layers including the bottleneck layer in the middle, and an output layer consisting of one node whose activation indicates the estimated level of interest. Unlike in bottleneck ASR systems, where the third hidden layer is only used during network training and not during decoding / feature generation, our networks for interest recognition employ *all* layers and thus perform dimensionality reduction and decorrelation *within* the network.

4. Features

4.1. Acoustic Features

The acoustic features applied in Section 5 correspond to the baseline feature set of the Interspeech 2010 Paralinguistic Challenge [3]. They are extracted via our real-time speech analysis toolbox openSMILE [11]. 1 582 acoustic features are obtained in total by systematic 'brute-force' feature generation in three steps: first, 38 low-level descriptors (see [3]) 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 statistical functionals are applied 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.

4.2. Linguistic Features

For linguistic feature extraction we apply our recently introduced multi-stream BLSTM-HMM ASR system [9] which was shown to prevail over conventional HMM systems in challenging spontaneous speech scenarios. 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 Section 3) into the speech decoding process.

In every time frame the HMM uses two independent observations: the MFCC features and a BLSTM phoneme prediction feature. The MFCC feature vector also serves as input for the BLSTM, which generates the maximum a posteriori phoneme estimate.

Via early fusion, we fuse linguistic information extracted by the BLSTM-HMM speech recognizer with the supra-segmental acoustic features described in Section 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. 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 TFs.

To reduce the size of the fused acoustic-linguistic feature space prior to subsequent dimensionality reduction and decorrelation within the bottleneck network, we conduct a cyclic Correlation based Feature Subset Selection (CFS) based on the TUM AVIC training set. As a result we obtain 92 selected acoustic features and combined acoustic-linguistic feature vectors of size 123, respectively.

Table 1: *Size of the hidden layers for networks with one hidden layer and bottleneck (BN) networks processing acoustic (Ac.) or combined acoustic-linguistic (Ac. + Ling.) information.*

classifier	**BN**	**size of hidden layers**	
		Ac.	**Ac. + Ling.**
BLSTM	yes	32-6-32	32-8-32
LSTM	yes	64-12-32	64-16-32
BRNN	yes	32-6-16	32-8-16
RNN	yes	64-12-16	64-16-16
BLSTM	no	32	32
LSTM	no	64	64
BRNN	no	16	16
RNN	no	32	32

5. Experiments and Results

We evaluated different neural network architectures with respect to their suitability for acoustic and acoustic-linguistic interest recognition: conventional recurrent neural networks, bidirectional recurrent neural networks (BRNN), LSTM networks, and bidirectional LSTM networks. For each network type we considered both, architectures with one hidden layer as used in [6] and [5], for example, and bottleneck structures consisting of three hidden layers (as discussed in Section 3). The number of memory blocks (or hidden nodes) per layer was optimized on the development set and can be seen in Table 1, e. g., the bottleneck-BLSTM processing acoustic and linguistic features applied 32 memory blocks in the first and third hidden layer and contained a bottleneck layer of size eight. Networks processing only acoustic features used slightly less memory blocks in the bottleneck layer (six for bidirectional networks). Note that simply increasing the number of hidden cells in networks consisting of one hidden layer or applying networks with an equal number of hidden cells (or memory blocks) in all three hidden layers led to lower performance on the development set than bottleneck architectures. The number of input nodes corresponds to the number of selected acoustic or combined acoustic-linguistic features. All memory blocks of the (B)LSTMs were composed of one memory cell. The networks had one (regression) output node whose activation represents the predicted level of interest.

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 multi-stream ASR system was parametrized as in [9]. 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).

Table 2 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 [3], 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

Table 2: *Results for interest recognition as defined in the Affect Sub-Challenge [3]: cross correlation obtained for different network architectures when using either acoustic (Ac.) or combined acoustic-linguistic (Ac. + Ling.) information with and without bottleneck (BN) structure; baseline results reported in [3] when applying unpruned REP-Trees with and without correlation-based feature selection (CFS); results reported in [12] and [13] when using SVM and GMM, respectively.*

classifier	CFS	BN	cross correlation	
			Ac.	Ac. + Ling.
BLSTM	yes	yes	0.459	**0.504**
LSTM	yes	yes	0.454	0.479
BRNN	yes	yes	0.427	0.440
RNN	yes	yes	0.434	0.433
BLSTM	yes	no	0.442	0.475
LSTM	yes	no	0.431	0.459
BRNN	yes	no	0.406	0.438
RNN	yes	no	0.422	0.439
REP-Trees	yes	-	0.439	0.435
REP-Trees [3]	no	-	0.421	0.423
SVM [12]	no	-	-	0.428
GMM [13]	no	-	0.390	-

an MLE of 0.148 (which is only 0.002 below the MLE reported in [3]) 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 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., [6]). Furthermore, bottleneck-(B)LSTM architectures consistently outperform networks with one hidden layer. The best performance can be obtained when applying bottleneck-BLSTM networks processing both, acoustic and linguistic features (CC of 0.504). 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 [3]) is shown in Table 2. The REP-Trees approach profits from feature selection via CFS but cannot compete with the bottleneck-BLSTM technique. Our results are notably better than the highest cross correlation that has ever been reported for the Affect Sub-Challenge so far (CC of 0.428 using Support Vector Machines (SVM) in combination with acoustic and linguistic information [12]) and prevail over the CC reported in [13] for Gaussian Mixture Models (GMM).

6. Conclusion

We proposed a speech-based framework for the assessment of a user's level of interest based on acoustic and linguistic information. Our approach exploits contextual knowledge via bidirectional Long Short-Term Memory networks which are able to model how the user's interest evolves over time. We combined the BLSTM technique with the idea of bottleneck nets by designing LSTM networks with multiple hidden layers, including a narrow (bottleneck) layer in the middle. This technique enables the generation of a compact low-dimensional feature representation within the network and led to improved interest recognition results. We showed that our bottleneck-BLSTM strategy achieves remarkable results on the Interspeech 2010 Paralinguistic Challenge task, outperforming all other methods which have been proposed for this task so far.

Future research should include multi-task learning of multiple emotional dimensions (e. g., valence, activation, and interest) as well as framewise modeling of interest as an alternative to applying turnwise statistical functionals as acoustic features.

7. Acknowledgements

The research leading to these results has received funding from the Federal Republic of Germany through the German Research Foundation (DFG) under grant no. SCHU 2508/4-1.

8. References

[1] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing Journal*, vol. 27, no. 12, pp. 1760–1774, 2009.

[2] S. Mota and R. Picard, "Automated posture analysis for detecting learner's interest level," in *Proc. of Workshop on CVPR for HCI*, Madison, 2003, pp. 49–55.

[3] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The Interspeech 2010 Paralinguistic Challenge," in *Proc. of Interspeech*, Makuhari, Japan, 2010, pp. 2794–2797.

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

[5] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling," in *Proc. of Interspeech*, Makuhari, Japan, 2010, pp. 2362–2365.

[6] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll, "Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 867–881, 2010.

[7] 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. IEEE Press, 2001, pp. 1–15.

[8] F. Grezl and P. Fousek, "Optimizing bottle-neck features for LVCSR," in *Proc. of ICASSP*, Las Vegas, NV, 2008, pp. 4729–4732.

[9] M. Wöllmer, F. Eyben, B. Schuller, and G. Rigoll, "A multi-stream ASR framework for BLSTM modeling of conversational speech," in *Proc. of ICASSP*, Prague, Czech Republic, 2011.

[10] A. Graves, "Supervised sequence labelling with recurrent neural networks," Ph.D. dissertation, Technische Universität München, 2008.

[11] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the Munich versatile and fast open-source audio feature extractor," in *Proc. of ACM Multimedia*, Firenze, Italy, 2010, pp. 1459–1462.

[12] J. H. Jeon, R. Xia, and Y. Liu, "Level of interest sensing in spoken dialog using multi-level fusion of acoustic and lexical evidence," in *Proc. of Interspeech*, Makuhari, Japan, 2010, pp. 2802–2805.

[13] R. Gajsek, J. Zibert, T. Justin, V. Struc, B. Vesnicer, and F. Mihelic, "Gender and Affect Recognition based on GMM and GMM-UBM modeling with relevance MAP estimation," in *Proc. of Interspeech*, Makuhari, Japan, 2010, pp. 2810–2813.