Speech-based non-prototypical affect recognition for child-robot interaction in reverberated environments

Martin Wöllmer, Felix Weninger, Stefan Steidl, Anton Batliner, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Wöllmer, Martin, Felix Weninger, Stefan Steidl, Anton Batliner, and Björn Schuller. 2011. "Speech-based non-prototypical affect recognition for child-robot interaction in reverberated environments." In *INTERSPEECH 2011 - 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, 3113–16. ISCA. https://www.isca-archive.org/interspeech_2011/wollmer11c_interspeech.html.

Speech-based Non-prototypical Affect Recognition for Child-Robot Interaction in Reverberated Environments

Martin Wöllmer[1], *Felix Weninger*[1], *Stefan Steidl*[2], *Anton Batliner*[2], *Björn Schuller*[1]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
[2]Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

woellmer@tum.de

Abstract

We present a study on the effect of reverberation on acoustic-linguistic recognition of non-prototypical emotions during child-robot interaction. Investigating the well-defined Interspeech 2009 Emotion Challenge task of recognizing negative emotions in children's speech, we focus on the impact of artificial and real reverberation conditions on the quality of linguistic features and on emotion recognition accuracy. To maintain acceptable recognition performance of both, spoken content and affective state, we consider matched and multi-condition training and apply our novel multi-stream automatic speech recognition system which outperforms conventional Hidden Markov Modeling. Depending on the acoustic condition, we obtain unweighted emotion recognition accuracies of between 65.4 % and 70.3 % applying our multi-stream system in combination with the SimpleLogistic algorithm for joint acoustic-linguistic analysis.

Index Terms: child-robot interaction, affective computing, acoustic-linguistic emotion recognition, reverberation

1. Introduction

Aiming to make human-machine communication more human-like, speech interfaces have emerged as a natural and easy-to-use input modality and are increasingly employed for a variety of applications including human-robot interaction, dialogue systems, voice command applications, virtual agents, and computer games. In addition to the spoken content, also paralinguistic and affective information can be automatically extracted from the speech signal. Thus, strategies towards automatic emotion recognition (AER) have attracted a lot of attention in recent years and are beginning to be used, e. g., for socially competent human-robot interaction [1].

While past research on AER has mostly been restricted to prototypical, acted, and speaker dependent emotion recognition, the focus of today's research is on speaker independence and on affective state estimation from non-prototypical, spontaneous speech as it is needed for real-life applications [2]. Reflecting these challenging conditions, which typically lead to recognition accuracies that are lower than those reported for prototypical emotions, the Interspeech 2009 Emotion Challenge [3] has been organized to define unified system training and test conditions involving spontaneous emotion recognition during child-robot interaction. Yet, one simplification of the Emotion Challenge task that might not necessarily hold for real-life systems is the restriction to speech captured by close-talk microphones. Thus, the effect of speech signal distortions caused by reverberation or background noise has been largely neglected in the Emotion Challenge – and generally in the field of speech-based emotion recognition. Only a few studies address the topic of noise robust AER, e. g., [4]. The impact of reverberation on AER from acoustic cues has been investigated in [5].

In this paper, we extend our recent research on affect recognition from reverberated speech [5] to systems that apply both, acoustic and linguistic features obtained via an automatic speech recognition (ASR) module. We examine how different microphones and room acoustics affect the quality of the ASR output on the one hand, and the accuracy of combined acoustic-linguistic emotion recognition on the other hand. To this end, we consider emotional child-robot interaction speech as contained in the FAU Aibo Emotion Corpus [6] in combination with different artificial and real reverberation conditions. Furthermore, we investigate matched, mismatched, and multi-condition training to increase the robustness of our proposed recognition engine. To further boost recognition performance and robustness, we employ our recently proposed multi-stream ASR system [7] which exploits context-sensitive phoneme estimates generated by a bidirectional Long Short-Term Memory (BLSTM) recurrent neural network [8]. The concept of BLSTM was shown to lead to enhanced ASR accuracies in challenging emotional speech scenarios [9, 10].

Our paper is structured as follows: Section 2 provides an overview over the FAU Aibo Emotion Corpus, Section 3 outlines the applied acoustic and linguistic features, and Section 4 briefly reviews the principle of our multi-stream ASR decoder. We describe our experiments in Section 5 before concluding in Section 6.

2. The FAU Aibo Emotion Corpus

The German FAU Aibo Emotion Corpus [6] with 8.9 hours of spontaneous, emotionally colored children's speech comprises recordings of 51 children at the age of 10 to 13 years from two different schools. Speech was transmitted with a wireless head set (UT 14/20 TP SHURE UHF-series with microphone WH20TQG) and recorded with a DAT-recorder. The sampling rate of the signals is 48 kHz; quantization is 16 bit. The data is downsampled to 16 kHz.

The children were given five different tasks where they had to direct Sony's dog-like robot Aibo to certain objects and through a given 'parcours'. The children were told that they could talk to Aibo the same way as to a real dog. However, Aibo was remote-controlled and followed a fixed, pre-determined course of actions, which was independent of what the child was actually saying. At certain positions Aibo disobeyed in order to elicit negative forms of emotions. The corpus is annotated by five human labelers on the word level using 11 emotion categories that have been chosen prior to the labeling process by it-

Table 1: *Segmental acoustic features: low-level descriptors (LLD) and functionals. For details, see [3].*

LLD ($16 \cdot 2$)	**Functionals** (12)
(Δ) ZCR	mean
(Δ) RMS Energy	standard deviation
(Δ) F0	kurtosis, skewness
(Δ) HNR	extremes: value, rel. position, range
(Δ) MFCC 1-12	linear regression: offset, slope, MSE

eratively inspecting the data. The units of analysis are not single words, but semantically and syntactically meaningful chunks (2.66 words per chunk on average, cf. [6]). Heuristic algorithms are used to map the decisions of the five human labelers on the word level onto a single emotion label for the whole chunk. The emotional states that can be observed in the corpus are rather non-prototypical, emotion-related states than 'pure' emotions. Mostly, they are characterized by low emotional intensity.

3. Feature Extraction

3.1. Acoustic Features

We extracted a set of 384 segmental acoustic features suited for static chunk-level classification, exactly corresponding to those used for the Interspeech 2009 Emotion Challenge baseline (Classifier Sub-Challenge) [2], including MFCC, prosodic, and voice quality features (Table 1). In fact, none of the Challenge participants could outperform the baseline features in the Feature Sub-Challenge [2].

3.2. Linguistic Features

To create linguistic features for early fusion with the chunk-level acoustic features, we converted the chunk-level ASR results, i. e., the reclassification of the training set, and the recognition of the test set, into a vector space representation by forming Bag-of-Words (BoW) vectors counting term frequencies. The components of the BoW vectors represent all words occurring in the reclassification of the training set by the ASR engine. As a result, the BoW feature space differs among training conditions. The BoW size ranges from 198 (training on room microphone data) to 379 (multi-condition training) since we intentionally do not use the ground truth transcriptions available in the FAU Aibo Emotion Corpus for building linguistic features, both to enforce realism, and to adapt to typical ASR confusions in the varying acoustic conditions.

4. Multi-Stream BLSTM-HMM

We implemented and evaluated two different ASR systems for linguistic feature generation: a standard single-stream Hidden Markov Model (HMM) system applying cross-word triphone acoustic models (see Section 5.3) and the multi-stream BLSTM-HMM system introduced in [7]. The multi-stream approach has shown enhanced recognition performance in challenging ASR conditions involving spontaneous, emotionally colored, and noisy speech. Our multi-stream decoder simultaneously models continuous MFCC observations and discrete context-sensitive phoneme estimates generated by a bidirectional Long Short-Term Memory recurrent neural network as two independent data streams. Details on the system architecture can be found in [7].

5. Experiments and Results

5.1. Interspeech 2009 Emotion Challenge Task

Along the lines of the Interspeech 2009 Emotion Challenge [3], the complete corpus is used for the experiments reported in this paper (i. e., not just chunks containing prototypical emotions). Yet, due to technical problems with the video camera recording the reverberated 'room microphone' data (see Section 5.2), only 17 076 of the 18 216 chunks could be used. The training set comprised 9 190 chunks and the test set consisted of 7 886 chunks. We considered the 2-class problem with the two main classes *negative valence* (NEG) and the default state *idle* (IDL, i. e. neutral) as defined for the Interspeech 2009 Emotion Challenge. A summary of this challenge is given in [2].

As the children of one school were used for training and the children of the other school for testing, the partitions feature speaker independence, which is needed in most real-life settings, but can have a considerable impact on classification accuracy. Furthermore, this partitioning provides realistic differences between the training and test data on the acoustic level due to the different room characteristics (see Section 5.2). Finally, it ensures that the classification process cannot adapt to socio-linguistic or other specific behavioral cues. Note that – as it is typical for realistic data – the two emotion classes are highly unbalanced (5 642 NEG-chunks vs. 11 434 IDL-chunks).

5.2. Acoustic Conditions

The data which was used for the Interspeech 2009 Emotion Challenge was recorded with a close-talk microphone (see Section 2) and will be called 'close-talk' (CT) in the following. Additionally, during creation of the FAU Aibo Emotion Corpus, the experiment was filmed with a video camera for documentary purposes. The child was not facing the microphone, and the camera was approximately 3 m away from the child. Thus, the audio channel of the videos is reverberated and contains background noises, e. g., the noise of Aibo's movements. While the recordings for the training set took place in a normal, rather reverberant class room, the recording room for the test set was a recreation room, equipped with curtains and carpets, i. e., with more favorable acoustic conditions. Thus, the data set provides realistic differences between training and test data on the acoustic level. This version will be called 'room microphone' (RM).

Another version [11] of the corpus was created using *artificial* reverberation: The data of the close-talk version was convolved with 12 different impulse responses recorded in a different room using multiple speaker positions (four positions arranged equidistantly on one of three concentric circles with the radii $r \in \{60\,\mathrm{cm}, 120\,\mathrm{cm}, 240\,\mathrm{cm}\}$) and alternating echo durations $T_{60} \in \{250\,\mathrm{ms}, 400\,\mathrm{ms}\}$ spanning 180°. The training and test set were evenly split in twelve parts, of which each was reverberated with a different impulse response, to enforce a roughly equal distribution of the impulse responses among the training and test set instances. This version will be called 'close-talk reverberated' (CTRV).

5.3. ASR Configuration and Training

The acoustic feature vectors processed by the ASR system consisted of cepstral mean normalized MFCC coefficients 1 to 12, log. energy, as well as first and second order delta coefficients. The framewise BLSTM phoneme predictor of the multi-stream system was trained on forced aligned (framewise) phoneme targets of the FAU Aibo Emotion Corpus training set. According

Table 2: *Single-stream HMM: ASR word accuracies for different training and test conditions. The best result per test condition is highlighted.*

word accuracy [%]	**test condition**			
training condition	CT	CTRV	RM	**mean**
CT	**85.28**	79.21	28.66	64.38
CTRV	82.86	**82.03**	48.82	71.24
RM	13.35	33.78	53.00	33.38
CT + CTRV + RM	83.05	81.11	**61.21**	**75.12**

Table 3: *Multi-stream BLSTM-HMM: ASR word accuracies for different training and test conditions. The best result per test condition is highlighted.*

word accuracy [%]	**test condition**			
training condition	CT	CTRV	RM	**mean**
CT	**87.03**	80.48	43.97	70.49
CTRV	85.33	**84.52**	56.83	75.56
RM	25.77	49.79	57.82	44.46
CT + CTRV + RM	83.76	82.13	**63.90**	**76.60**

to our past experience [7], we chose three hidden layers of size 56, 150, and 56, respectively, to model 53 German phonemes as well as *silence*, *short pause*, and *non-verbal events*. All other parameters of the multi-stream ASR system, such as the stream weight of the BLSTM phoneme prediction feature stream, were configured as in [7].

The underlying HMM system applied phoneme models consisting of three emitting states (left-to-right HMMs) with eight Gaussian mixtures. Initial monophones HMMs were mapped to tied-state cross-word triphone models with shared state transition probabilities. Both, acoustic models and a back-off bigram language model were trained on the training set of the FAU Aibo Emotion Corpus.

Table 4: *Unweighted accuracies (UA) for acoustic, linguistic, and combined acoustic-linguistic classification of the test set by feature-level fusion with BoW vectors from the single-stream HMM speech recognizer. The best result per test condition is highlighted.*

UA [%]	**test condition**			
training condition	CT	CTRV	RM	**mean**
acoustic				
CT	67.90	53.99	59.83	60.57
CTRV	59.97	67.22	60.27	62.48
RM	66.32	63.03	64.96	64.77
CT + CTRV + RM	68.20	66.24	60.40	64.95
linguistic (single-stream HMM)				
CT	64.76	64.92	54.67	61.45
CTRV	63.59	63.15	58.05	61.59
RM	55.47	58.06	60.20	57.91
CT + CTRV + RM	63.38	62.99	60.29	62.22
acoustic + linguistic (single-stream HMM)				
CT	**70.08**	59.27	60.94	63.43
CTRV	60.28	**68.55**	62.44	63.76
RM	65.86	63.58	**65.41**	64.95
CT + CTRV + RM	68.92	67.96	62.48	**66.46**

5.4. ASR Results

Tables 2 and 3 show the word accuracies (WA) when applying standard triphone acoustic models and the multi-stream BLSTM-HMM approach, respectively. We consider four different ASR training conditions: training on data recorded by the close-talk microphone (CT), artificially reverberated data (CTRV), data recorded by the room microphone (RM), and all data (CT + CTRV + RM). Accuracies are consistently higher for the multi-stream model with performance gains of up to 16 % (absolute) when training on RM data and testing on CTRV data. This indicates that BLSTM context modeling within the multi-stream technique leads to higher robustness with respect to different reverberation conditions. However, also for 'friendly' scenarios, e. g, training and testing on data recorded by close-talk microphones, the multi-stream model prevails over standard HMMs (word accuracy of 87.03 % vs. 85.28 %). These accuracies are notably higher than those reported in [12], for example. As expected, matched condition training performs best, with the exception that RM data is best recognized using models trained on data reflecting all three acoustic conditions. Generally, multi-condition training leads to high accuracies for all test conditions and achieves the best average ASR performance (WA of 76.6 % for the multi-stream model).

5.5. AER Classification Strategy

To investigate the impact of ASR performance on emotion recognition, we evaluated linguistic and joint acoustic-linguistic analysis by early feature-level fusion using the SimpleLogistic algorithm [13] implemented in the Weka toolkit [14]. It is based on boosting of one-dimensional regression functions, thereby implicitly performing a feature relevance analysis and selection. This technique seems to be particularly suited for feature-level fusion dealing with varying reliability of features according to acoustic conditions. The number of boosting iterations was cross-validated on the training set, using the default parameters in the Weka toolkit for straightforward reproducibility. Since the class distribution in the training set of the FAU Aibo Emotion Corpus is heavily unbalanced, we applied the Synthetic Minority Oversampling Technique (SMOTE).

Table 5: *Unweighted accuracies (UA) for linguistic and acoustic-linguistic classification of the test set by feature-level fusion with BoW vectors from the multi-stream BLSTM-HMM speech recognizer. The best result per test condition is highlighted.*

UA [%]	**test condition**			
training condition	CT	CTRV	RM	**mean**
linguistic (multi-stream HMM)				
CT	65.21	64.53	56.54	62.10
CTRV	63.90	63.58	58.74	62.07
RM	56.44	59.96	60.64	59.01
CT + CTRV + RM	64.07	63.28	60.44	62.60
acoustic + linguistic (multi-stream HMM)				
CT	**70.32**	59.34	62.19	63.95
CTRV	60.34	**68.61**	63.05	64.00
RM	65.80	64.05	**65.43**	65.09
CT + CTRV + RM	69.16	67.84	62.96	**66.65**

5.6. AER Results and Discussion

In Table 4, we present the unweighted accuracies (UA) for emotion recognition by BoW linguistic features obtained from single-stream HMM ASR, both with and without acoustic features. For reference, we also show the results by acoustic features only. For CT, CTRV, and multi-condition training, these are similar to the ones obtained by Support Vector Machines (SVM) in [5]; for RM training however, the SimpleLogistic classifier yields a significant ($p < 0.005$) performance gain over SVM in the CT (66.32 vs. 61.61 % UA) and RM (64.96 vs. 62.72 % UA) test cases. Best average performance is achieved by multi-condition training (64.95 % UA).

Furthermore, linguistic features on their own deliver remarkable performance: When using ASR features from CT data for training, 64.76 % and 64.92 % UA are achieved in the CT and CTRV test conditions, respectively. Overall, a strong correlation with the word accuracies from Table 2 can be seen, with multi-condition training showing best average performance (62.22 % UA) once more.

Finally, by fusion of acoustic and linguistic information a significant ($p < 0.005$) performance improvement over acoustic features, from 67.90 % to 70.08 % UA is observed for matched condition CT training and testing. While in RM testing, the clean acoustic-linguistic classifier prevails over both pure acoustic and linguistic analysis (60.94 % UA vs. 59.83 % and 54.67 %, respectively), this is not the case for CTRV testing, where a drop in performance (59.27 % vs. 64.92 % UA) compared to linguistic features is observed, which is arguably caused by the poor performance of acoustic features in that particular setup (53.99 % UA). Remarkably, on average over all test conditions, fused acoustic-linguistic analysis using multi-condition training (66.46 % UA) considerably outperforms linguistic (62.22 %) and acoustic analysis (64.95 % UA). The best performance on RM, i. e., realistically reverberated, data is obtained by fused acoustic-linguistic analysis trained on RM (65.41 % UA) – note that this is not matched condition training in a strict sense, since the training and test set were recorded in different acoustic settings (see Section 5.2). This suggests that whenever the acoustic conditions that the emotion classifier has to face are known to a certain degree (corresponding to CT and CTRV testing), multi-condition training is most promising; for unknown conditions (RM testing), training on realistically reverberated data is to be preferred, even if that data does not exactly match the acoustic conditions to be faced.

Table 5 shows the results for linguistic and acoustic-linguistic AER when applying the multi-stream BLSTM-HMM speech recognizer for linguistic feature generation. For almost all training and test conditions, we observe higher accuracies than for the recognition engine using conventional HMM ASR. Trends are similar to those in Table 4, i. e., matched condition training performs best while multi-condition training leads to the best average accuracy.

6. Conclusion

We analyzed the effect of reverberation on automatic speech and emotion recognition in a child-robot interaction scenario involving spontaneous speech and non-prototypical emotions. Reverberation tends to degrade acoustic, linguistic, and combined acoustic-linguistic emotion recognition performance, however, the usage of reverberated training material can largely compensate the decrease of both, speech and emotion recognition accuracy. Multi-condition training leads to good performance for all reverberation conditions and reaches accuracies comparable to matched condition training. This shows that including reverberated data in the training set leads to more robust models – even if the training conditions do not exactly match the acoustic conditions during testing. Applying a multi-stream BLSTM-HMM ASR system, acoustic-linguistic AER accuracies of up to 70.3 % can be obtained for the recognition of negative emotions, which corresponds to results that were previously only reported for the fusion of multiple recognition engines [12].

Future research should focus on the combination of multi-condition training and speech feature enhancement.

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 nos. SCHU 2508/4-1 and SCHU 2508/2-1.

8. References

[1] A. Delaborde and L. Devillers, "Use of nonverbal speech cues in social interaction between human and robot: emotional and interactional markers," in *Proc. of AFFINE*, Firenze, Italy, 2010, pp. 75–80.

[2] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication, Special Issue on "Sensing Emotion and Affect – Facing Realism in Speech Processing"*, 2011, to appear.

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

[4] A. Tawari and M. Trivedi, "Speech emotion analysis in noisy real world environment," in *Proc. of ICPR*, Istanbul, Turkey, 2010, pp. 4605–4608.

[5] F. Weninger, B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognition of Nonprototypical Emotions in Reverberated and Noisy Speech by Nonnegative Matrix Factorization," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, 2011, article ID 838790, 16 pages.

[6] S. Steidl, *Automatic Classification of Emotion-Related User States in Spontaneous Speech*, Logos, Berlin, Germany, 2009.

[7] 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.

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

[9] M. Wöllmer, F. Eyben, A. Graves, B. Schuller, and G. Rigoll, "Bidirectional LSTM networks for context-sensitive keyword detection in a cognitive virtual agent framework," *Cognitive Computation*, vol. 2, no. 3, pp. 180–190, 2010.

[10] 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.

[11] A. Maier, C. Hacker, S. Steidl, E. Nöth, and H. Niemann, "Robust Parallel Speech Recognition in Multiple Energy Bands," in *Proc. of Pattern Recognition, DAGM Symposium*, Vienna, Austria, 2005, pp. 133–140.

[12] B. Schuller, F. Metze, S. Steidl, A. Batliner, F. Eyben, and T. Polzehl, "Late fusion of individual engines for improved recognition of negative emotion in speech – learning vs. democratic vote," in *Proc. of ICASSP*, Dallas, Texas, 2010, pp. 5230–5233.

[13] N. Landwehr, M. Hall, and E. Frank, "Logistic Model Trees," *Machine Learning*, pp. 161–205, May 2005.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, 2009.