Emotional and User-Specific Acoustic Cues for Improved Analysis of Naturalistic Interactions

Ingo Siegert¹

Published online: 16 September 2015
© Springer-Verlag Berlin Heidelberg 2015

1 Introduction

Human–computer interaction recently received increased attention. Besides making the operation of technical systems as simple as possible, one goal is to enable a natural interaction. However, today’s speech-based operation still seems artificial: only the content of speech is evaluated. The way in which something is said remains unconsidered, although it is well-known that emotions are important to communicate successfully, too. “Companion systems” aim to fill this gap by adapting to the user’s individual skills, preferences and emotions [5]. The dissertation discussed in the following presents methodological improvements on speech-based emotion recognition and interpretation [2].

2 Automatic Emotion Recognition—Challenges

Automatic speech-based emotion recognition is a branch of data-driven pattern recognition. Insights are gathered from samples, as it is difficult to rely on empirical evidences of emotion psychology: there is no uniform emotion representation, and only a rather descriptive definition of appropriate emotion-distinctive features.

Initially, automatic emotion recognition was—due to the lack of suitable data sets—usually based on acted and very expressive emotional utterances. In this case, very good recognition rates of over 80 % distinguishing up to seven emotions can be achieved. However, for human–computer

interactions, these recognisers were unsuitable because naturalistic emotions are distinctively less expressive. Therefore, in collaboration with psychologists, naturalistic interaction scenarios were developed to collect relevant data sets with participants from different groups, who were not given specifications for “acting”. This led to decreased recognition rates of only 60 % on this type of data [1].

To increase the recognition results and improve the analysis of speech-based naturalistic interactions, four issues were investigated in the thesis [2] which are explained in the remainder of the paper. In particular, it was examined if further technically observable cues improve the emotion recognition and interaction control in naturalistic human computer interaction.

3 Investigated Open Issues

The first open issue dealt with the generation of a reliable assignment of emotion labels. Since in natural interactions emotional reactions are not specified, a reliable class assignment has to be created after the recordings by a suitable annotation. In the thesis, it was shown that, for a naturalistic human–computer interaction, the annotation quality and reliability can be increased by combining audio and video data as well as using contextual information. These methodological improvements ensure an annotation of high quality [3].

The second open issue examined whether the speakers’ gender or age-group have to be considered for the improvement of emotion recognition. The vocal tract differs between male and female speakers and is also changing due to aging. This also affects the acoustic features used for emotion recognition. Through experiments with different datasets it was shown that the recognition

✉ Ingo Siegert
ingo.siegert@ovgu.de

¹ Cognitive Systems Group, FEIT-IKT, Otto von Guericke University Magdeburg, Magdeburg 39106, Germany

performance is significantly improving when considering gender or age. In some cases a combination of both factors achieved an even further improvement [4]. Subsequently, the gender and age-group specific modelling was extended to the fusion of continuous, fragmentary, audio–visual data. It was shown that an improvement in the fused recognition is possible, even when the speech data is not available for the entire data stream.

The third open issue expanded the object of investigation to naturalistic human–computer interactions and examines whether certain acoustic feedback signals can be used for an evaluation of emotions. The work focused on discourse particles, such as “hm” or “h”, which are short vocalizations, interrupting the flow of speech. As they are semantically meaningless, only their intonation is relevant. First, the thesis revealed that they serve as an indicator of difficult parts of the interaction: in challenging dialogues significantly more discourse particles are uttered than in simple dialogues. A further special feature of discourse particles is that they have specific functions in a dialogue depending on their intonation. In the thesis, it was shown that the most common function “thinking” is robustly distinguishable from all other dialogue functions by using the intonation only.

The fourth open issue dealt with the modelling of temporal evolution of emotions. The system’s reaction should not be based only on a single observation of emotions or interaction patterns of the user. Instead his long-term emotional development has to be considered, including personal factors. For this purpose, a mood-model was introduced, calculating the mood from the course of observed emotions, based on a physically inspired spring model. Furthermore, the integration of the user’s personality trait “extraversion” allows an individual adaptation of the model.

4 Conclusion

The dissertation extends the pure acoustic emotion recognition by considering further modalities, speaker characteristics, feedback signals and personality traits. This enables technical systems to examine longer-lasting natural interactions and dialogues and to identify critical situations. Thereby, the human–computer interaction is conducted more naturalistic.

Technical systems that use this extended emotion recognition adapt to their users and thus become his attendant and ultimately his companion.

Acknowledgments The work presented in this paper was done within the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” (www.sfb-trr-62.de) funded by the German Research Foundation (DFG).

References

1. Schuller B, Batliner A, Steidl S, Seppi D (2011) Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. *Speech Commun* 53:1062–1087
2. Siegert I (2015) Emotional and user-specific cues for improved analysis of naturalistic interactions. Ph.D. thesis, Otto von Guericke University, Magdeburg
3. Siegert I, Böck R, Wendemuth A (2014) Inter-rater reliability for emotion annotation in human–computer interaction—comparison and methodological improvements. *J Multimodal User Interfaces* 8:17–28
4. Siegert I, Philippou-Hübner D, Hartmann K, Böck R, Wendemuth A (2014) Investigation of speaker group-dependent modelling for recognition of affective states from speech. *Cogn Comput* 6(4):892–913
5. Wendemuth A, Biundo S (2012) A companion technology for cognitive technical systems. In: Esposito A, Esposito AM, Vinciarelli A, Hoffmann R, Müller V (eds) *Cognitive behavioural systems, LNCS*, vol 7403. Springer, Berlin, pp 89–103

Ingo Siegert is postdoctoral associate in the Cognitive Systems Group at the Otto von Guericke University Magdeburg and researcher in the SFB/TRR 62 “A Companion-Technology for Cognitive Technical Systems”. In his research he focuses on naturalistic human–computer interaction in terms of emotion annotation, incorporation of speaker groupings for improved recognition, and utilisation of further interaction patterns, as filled pauses or discourse particles, to improve the evaluation of interactions. Furthermore, he presented a modelling technique for the temporal evolution of emotions.