ACOUSTIC ADDRESSEE-DETECTION – ANALYSING THE IMPACT OF AGE, GENDER AND TECHNICAL KNOWLEDGE

Ingo Siegert, Tang Shuran, Alicia Flores Lotz
Institute for Information and Communications Engineering, Cognitive Systems Group,
Otto-von-Guericke University, 39016 Magdeburg, Germany, www.cogsy.de

Introduction Today, in technical dialog-systems quite multiple solutions are implemented to detect if a system should react to an uttered speech command. Typically used solutions are push-to-talk and keywords. Unfortunately, these solutions constitute an unnatural interaction to overcome the problem that the system is not able to detect when it is addressed. Besides these issues, the actual preferred keyword method can result in confusions when the keyword has been said but no interaction with the system was intended by the user. Therefore, technical systems should be able to perform an addressee-detection. Various aspects have already been investigated in this field of research, however most of them pursue a multimodal approach including textual and/or visual information achieving up to 93% unweighted average recall. In our research, we limit ourselves to the pure acoustic information, as we assume that humans are talking differently to technical systems and to humans.
Considering speakers of different age-, sex- and technical background-groups, we analysed how a technical system and another human being is being addressed in two separate setups. An acoustics-based addressee-detection system was utilized and it was investigated to which extent the different speaker groups influence the recognition rate in inter- and intra-group experiments.

Utilized Data For our study we utilize the LAST MINUTE Corpus. It contains 130 high-quality multi-modal recordings of German speaking subjects obtained from Wizard of Oz experiments. This part is referred to as Human-Computer-Interaction (HCI)-part. Furthermore, 73 of these subjects underwent a semi-structured interview, which followed the HCI-part experiment. This part is referred as Human-Human-Interaction (HHI)-part. A sub-set of 24 speakers having high quality recordings under the same acoustic conditions available for both the HCI and the HHI-part,serves as data base for the presented experiments. The addressee-detection problem conducted in this setup is to detect whether an utterance originates from the HCI- or the HHI-part.

Recognition Results For feature extraction, we used the emobase set from OpenSMILE, as a good compromise between feature size and feature accuracy, and we applied standardization as normalisation method. As classifier a linear Support-Vector-Machine was used. We achieved a mean F-measure of 0.9218 with a standard deviation of 0.01564 using all speakers in a leave-one-speaker-out validation scheme. This shows that the same level of accuracy of multimodal information processing can be obtained from speech only. The recognition results regarding the inter- and intra-group experiments will be presented in detail in the full paper. But we can already state that the results don't differ significantly between the different experiments.

Conclusion As the recognition results for the different groups are similar, it can be assumed that there is a general way of communicating with technical systems which can be retrieved by speech analysis alone, especially when the technical system has a very technical sounding voice and the participant is talking either to a machine or to another human being. The result of this paper will serve as basis for consecutive studies analysing the influence of certain factors (human-likeness of the technical system, addressing both machine and human being simultaneously, presence of the technical system).