ACOUSTIC ADDRESSEE-DETECTION – ANALYSING THE IMPACT OF AGE, SEX 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
ingo.siegert@ovgu.de

Abstract: Today, in technical dialog-systems diverse 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. Moreover, 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 than 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 on two subsets. An addressee detection system based on acoustics-based was utilized and it was investigated to which extent the different speaker groups influence the recognition rate in inter- and intra-group experiments. Our approach achieves competitive results of 84.45% to 98.06% unweighted average recall and 88.35% to 95.63% F_1 score.

1 Introduction

Human-computer interaction (HCI) receives continued attention, the market for commercial voice assistants is rapidly growing. Besides making the operation of technical systems as simple as possible, voice assistants should enable a natural interaction. Therefore, one aspect that still needs improvement is to automatically recognise the addressee of a user's utterance.

Diverse solutions are implemented to detect if a system should react to an uttered speech command. A short literature research is given in Section 2. 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. Especially, 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.

By way of contrast, we limit ourselves to the pure acoustic information, as actual commercial voice assistant systems do not support video (gaze) analyses. We hereby assume that the speaking style in addressing technical systems is a general pattern regardless of age, sex, or technical affinity. To prove this assumption, we designed several inter- and intra-group addressee detection experiments.

Due to the lack of proper databases providing, on the one hand, speech data of users talking to both technical systems and other humans and, on the other hand, additional information about these speaker's age, sex and technical affinity, we made a compromise and used the LAST MINUTE corpus. Details about this dataset are given in Section 3. The data preparation, to reduce the effect of unwanted influences. is explained in Section 4.

For our experiments, we used state of the art classification methods and analysed both the completeness of recognition results as well as the usefulness of the recognition results. The experimental design is specified in Section 5. The results are given in Section 6. Finally, Section 7 concludes the paper and provides an outlook an further research.

2 State of the Art

Most of the addressee detection studies for speech enabled systems utilize self-recorded databases either with one human and a technical system or groups of humans (mostly two) interacting with each other and a technical system [1, 2, 3, 4, 5] or teams of robots and teams of humans [6]. These studies are mostly done using one specific scenario, just a few researchers analyse how people interact with technical systems in different scenarios [7, 8]. In these studies, the technical system is either a robot [6, 9], a research system [1, 2], or a Wizard-of-Oz (WOZ)-experiment [5].

Most authors use either eye-gaze, or language related features (utterance length, keyword, trigram-model), or a combination of both [2, 6, 7, 9, 5]. Regarding the experimental results on acoustic data, researchers report different measures. One common measure is the Equal Error Rate (EER) in combination with the Detection Error Tradeoff (DET). In [3], 150 multiparty interactions of 2 to 3 people playing a trivia question game with a computer are utilized. The dataset comprises audio, video, beamforming, system state and ASR information. For acoustic analyses, energy, energy change and temporal shape of speech contour features, in total 47, are used to train an adaboost classifier. They achieved an EER of 13.88%.

In [1], data of 38 sessions of two people interaction with a "Conversational Browser" is used. Using energy and speaking rate features as well as energy contour features to train a Gaussian Mixture Model (GMM) together with linear logistic regression and boosting, the authors achieved an EER of 12.63%. The same data is used in [4]. Their best acoustic EER of 12.5% is achieved using a GMM with adaptive boosting of energy contour features, voice quality features, tilt features, and voicing onset/offset delta features. The authors of [8] used two different experimental settings (standing and sitting) of a WOZ data collection with 10 times two speakers interacting with an animated character. They employed a Support Vector Machine (SVM) and four supra-segmental speech features (F_0, intensity, speech rate and duration) as well as two speech features describing the difference from all speakers average for F_0 and intensity. The reported acoustic accuracy is 75.3% for the participants standing and 80.7% for the participants sitting. A recent study utilizing a public available well-known corpus (Smart Web SVC data) achieves up to 82.2% Unweighted Average Recall (UAR) using the IS13_ComParE feature set (reduced to 1000 features using feature selection) with an SVM [10]. The drawback of this corpus is, that it actually only provides offtalk data. Thus the experiments are analysing offtalk vs. non-offtalk rather than an addressee detection problem.

3 Utilized Data

For our study we utilize the LAST MINUTE Corpus (LMC) [11]. It contains 130 high-quality multi-modal recordings of German speaking subjects obtained from WOZ experiments collected in 2010/2011. It is already the object of examination regarding affective state recognition

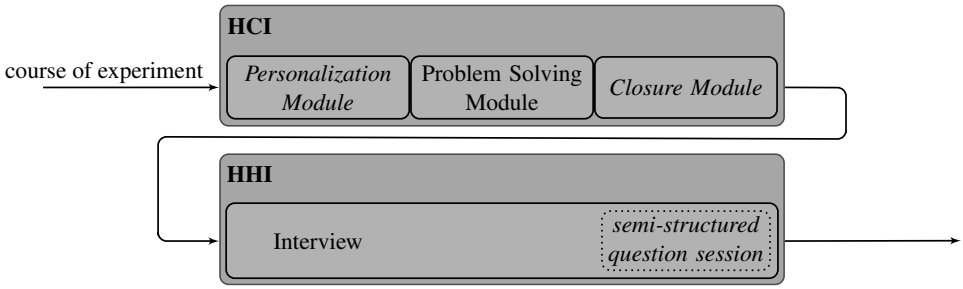

Figure 1 – A sketch of the experimental procedure. The considered participants first conducted the HCI-part consisting of three different modules. Afterwards they underwent an interview with a human partner. Both parts have a semi-structured section, denoted in *italics*.

[12, 13] and linguistic turns [14]. LMC is referred to as HCI-part in this paper. Additionally, 73 subjects underwent a semi-structured interview, subsequently to the interaction with the technical system. This interview is referred as human-human interaction (HHI)-part in this paper. The course of the experiment is given in Figure 1.

Furthermore, the corpus was recorded with several opposing speaker groups, young vs. elderly speakers and male vs. female speakers. The younger group was represented by subjects being 18 to 28 years old. The elder group consists of subjects being older than 60 years. By evaluating the TA-EG questionnaire [15], we could also identify two groups of technical affinity. By using a median split we identified participants with low technical affinity and participants with high technical affinity. We suppose that people of higher technical affinity are more used to working with technical machines and thus have less reluctance in speaking to machines. The distribution of speakers for each group (age, sex, technical affinity) is nearly balanced, see Table 1a.

Although this dataset does not provide speech data of speakers talking concurrently with technical systems and with human beings, it is an acceptable compromise. It provides an adequate corpus size and additional user characteristics.

The **HCI-part** is distinguished into three modules, with two different dialog styles [16, 17]. The *personalization module* and the *closure module*, represent a semi-structured interview. They have the purpose of making the subject familiar with the system and ensuring natural behavior. The *problem solving module* revolves around an imaginary journey to the unknown place Waiuku. The subjects have to prepare the journey, by packing the suitcase, and select clothing and other equipment by using voice commands. This module has a command-like regularized dialog style.

The **HHI-part** is conducted as an interview with a human being. The subjects were asked to described their individual experience of the experimental interaction and the simulated system [18]. The interview focused in particular on the subjects‘ emotions occurring during the interaction, the subjects‘ subjective ascriptions to the system and the subjects’ overall evaluation of the system. In the end of the interview, more structured questions were asked, so that this part is comparable to the semi-structured interview modules (personalization and closure) of the HCI-part.

4 Data Preparation

To reduce the effect of unwanted influences, we put a lot of effort in data-preparation. First of all, we selected the interactions of 24 speakers providing high quality recordings under the same acoustic conditions available for both the HCI and the HHI-part. Afterwards, a manual cleaning stage of these utterances was conducted to remove all samples having background noise,

laughter, or overlaps with the system output/interviewer. This sub-set is denoted as "set_{24}" in the following. In this context, it was taken into account that the considered utterances for the HCI and HHI part are about the same length. Especially longer speaking parts in the interview were neglected.

To furthermore meet the criticism that the HCI and HHI part represent different domains or tasks, we selected the semi-structured module of the HCI-part and the semi-structured question session from the HHI-part to gain data from the same type of task. This sub-set is denoted as "$set_{24_{structured}}$" in the following. The number of samples and duration for both classes (HCI and HHI) and for both sub-sets are given in Table 1b.

Group	Male	Female	Total
Young	4 (3/1)	9 (4/5)	13(7/6)
Elderly	7(4/3)	4(1/3)	11(5/6)
Total	11(7/4)	13(5/8)	11(12/12)

(a) Speaker groups in the considered sub-set of the LAST MINUTE corpus. Distribution of technical affinity is given in braces: (low technical affinity vs. high technical affinity).

Set	Samples		Duration	
	HCI	HHI	HCI	HHI
set_{24}	1644	8774	62.9 min	188,8 min
$set_{24_{structured}}$	418	6112	32.6 min	132.1 min

(b) Number of samples and duration in minutes for the considered sub-sets of the LAST MINUTE corpus.

Table 1 – Overview of speaker groupings and dataset characteristics for the considered sub-sets of the LAST MINUTE corpus.

5 Experimental Setup

The addressee detection problem conducted in this setup is to detect whether an utterance originates from the HCI- or the HHI-part. For this, we conducted state-of-the-art recognition experiments comparable to [19] for our addressee detection problem.

For feature extraction, we used the "emobase" feature set provided by openSMILE [20] as a good compromise between feature size and feature accuracy. It comprises 988 features derived from 19 functionals calculated for 52 Low-Level-Descriptors (LLDs) and has been successfully used for various classification experiments [21, 22]. Differences between the data samples of different speakers were then eliminated using standardization [23].

As recognition system, a SVM with linear kernel and a cost factor of 1 was utilized with WEKA [24]. We applied a Leave-One-Speaker-Out (LOSO) validation and calculated the Unweighted Average Recall (UAR), Unweighted Average Precision (UAP) and F_1-score (F-measure) for each validation step. This strategy allows us to revise the generalisation ability of the actual experiment. Finally, the $\overline{\text{UAR}}$, $\overline{\text{UAP}}$ and $\overline{\text{F-measure}}$ were calculated as the average over all speakers.

6 Recognition Results

To test our assumption that the speaking style in addressing technical systems is a general pattern regardless of age, sex, or technical affinity, we conducted several inter- and intra-group experiments. For both experiments we used a LOSO verification strategy. Thereby, it is secured that test speaker samples are not used for the classifier training. For inter-group experiments, the classifier is trained with speakers of the same group, e.g. other male speakers for male

speaker testing. For intra-group experiments, the test speaker is from another group than the train speakers, e.g. a female test speaker while training with male speakers. For both types of experiments the experiment runs are repeated for all speakers of the test group. Furthermore, these experiments are conducted for both sub-sets set_{24} and $set_{24_{structured}}$. The results are given in Table 2a and Table 2b, respectively.

Train set	Test set	$\overline{UAR}$	$\overline{UAP}$	$\overline{F\text{-measure}}$
		All		
All	All	93.67	93.98	92.18
		Sex		
Male	Male	94.95	90.39	92.61
Female	Male	94.90	93.65	92.90
Male	Female	89.87	92.16	91.00
Female	Male	94.97	93.35	93.07
		Age		
Young	Young	96.48	96.12	95.63
Elderly	Elderly	93.08	94.90	93.98
Young	Elderly	98.06	88.92	93.27
Elderly	Young	96.13	89.31	92.59
		Technical Affinity		
Low	Low	89.27	93.11	91.15
High	High	90.93	91.55	91.24
Low	High	91.39	93.82	92.59
High	Low	92.96	94.53	93.71

(a) Sub-set: set_{24}

Train set	Test set	$\overline{UAR}$	$\overline{UAP}$	$\overline{F\text{-measure}}$
		All		
All	All	87.00	94.85	90.75
		Sex		
Male	Male	87.58	93.30	90.34
Female	Female	86.99	94.45	90.57
Male	Female	86.68	91.53	89.04
Female	Male	87.18	93.29	90.13
		Age		
Young	Young	87.49	95.45	91.30
Elderly	Elderly	84.35	94.40	89.09
Young	Elderly	89.76	86.97	88.34
Elderly	Young	87.49	96.09	89.56
		Technical Affinity		
Low	Low	87.05	95.89	90.25
High	High	88.10	93.70	90.81
Low	High	87.04	93.75	90.27
High	Low	86.37	96.57	91.19

(b) Sub-set: $set_{24_{structured}}$

Table 2 – $\overline{UAR_h}$, $\overline{UAP}$ and $\overline{F\text{-measure}}$ for the different combinations of inter- and intra-speaker groupings for both sub-sets.

Regarding Table 2a and Table 2b, it can be seen that the addressee detection experiments achieve high recall and precision values, independently of the type of experiment. The classification rates for intra- and inter-group experiments are outstanding. For the sub-set set_{24} we achieve an average UAR above 89.27% and an average UAP above 88.92%. For the sub-set $set_{24_{structured}}$ the recognition rates are slightly lower with an average UAR above 84.35% and an average UAP above 86.97%.

In comparison to classification results of other researchers (Section 2), we can state that the results on the $set_{24_{structured}}$ are competitive, although they are not directly comparable, due to other evaluation measures and other data sources.

7 Conclusion

In this paper, we analysed the speaking style in addressing technical systems using addressee detection experiments. We assume for addressing technical systems similar speaking styles are used regardless of the users' age, sex, or technical affinity. To prove this assumption, we designed several inter- and intra-group addressee detection experiments using data from the

LAST MINUTE corpus and subsequently conducted interviews with some of the participants. This dataset has the advantage that various socio-demographic characteristics are known. Besides age and sex this also includes psychometric data. In this paper we concentrated on the personality trait "technical affinity", as we suppose this is an important factor in interacting with technical systems.

To reduce possible side-effects, we selected samples of 24 speakers recorded under the same acoustic conditions in both parts (HCI and HHI). We further performed a data cleaning to remove all samples, which include other sounds than the speaker's voice. Besides using all remaining samples of the 24 speakers, we also selected samples gathered via semi-structured question sessions in both HCI and HHI parts. These samples can be considered as being in the same domain. Afterwards, state-of-the-art recognition experiments are conducted, using the emobase feature-set of OpenSMILE, a linear SVM and a LOSO validation scheme. We reported $\overline{\text{UAR}}$, $\overline{\text{UAP}}$ and $\overline{\text{F-measure}}$.

Restrictively to our experiments, it has to be noted that in our data, firstly the participants are either talking to a machine or to another human being in separated recording sessions. Secondly, the domains of the HCI-part and HHI-part are not identical, apart from the semi-structured question sessions. But, it is very unlikely that the domains for HCI and HHI are identical in general.

The achieved recognition results for the different inter- and intra-group experiments within both sub-sets are quite similar. Therefore, it can be assumed that there is a general way of communicating with technical systems which can be retrieved by speech analysis alone and which can be modeled using state-of-the-art classification methods and a suitable (large) feature set. The result of this paper will serve as basis for consecutive studies analysing the excact distinctive features and their attributing to acoustic characteristics as well as the influence of certain factors, amongst those are human-likeness of the technical system, addressing both machine and human being simultaneously, presence of the technical system.

Acknowledgment

One of us (A. Lotz) wishes to acknowledge funding from the European Union's Horizon 2020 research and innovation programme in the project "ADAS&Me" under grant agreement No. 68890.

References

[1] SHRIBERG, E., A. STOLCKE, D. HAKKANI-TÜR, and L. HECK: *Learning when to listen: Detecting system-addressed speech in human-human-computer dialog*. In *Proc. of the INTERSPEECH'12*, pp. 334–337. Portland, USA, 2012.

[2] VINYALS, O., D. BOHUS, and R. CARUANA: *Learning speaker, addressee and overlap detection models from multimodal streams*. In *Proc. of the 14th ACM ICMI*, pp. 417–424. 2012. doi:10.1145/2388676.2388770.

[3] TSAI, T., A. STOLCKE, and M. SLANEY: *Multimodal addressee detection in multiparty dialogue systems*. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2314–2318. 2015.

[4] SHRIBERG, E., A. STOLCKE, and S. RAVURI: *Addressee detection for dialog systems using temporal and spectral dimensions of speaking style*. In *Proc. of the INTERSPEECH'13*, pp. 2559–2563. Lyon, France, 2013.

[5] VAN TURNHOUT, K., J. TERKEN, I. BAKX, and B. EGGEN: *Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features*. In *Proc. of the 7th ACM ICMI*, pp. 175–182. 2005. doi:10.1145/1088463.1088495.

[6] DOWDING, J., W. J. CLANCEY, and J. GRAHAM: *Are you talking to me? dialogue systems supporting mixed teams of humans and robots*. In *AIAA Fall Symposium Annually Informed Performance: Integrating Machine Listing and Auditory Presentation in Robotic Systems*. Washington, DC; United States, 2006.

[7] LEE, H., A. STOLCKE, and E. SHRIBERG: *Using out-of-domain data for lexical addressee detection in human-human-computer dialog*. In *Proc. NAACL*, pp. 221–229. Atlanta, USA, 2013.

[8] BABA, N., H.-H. HUANG, and Y. I. NAKANO: *Addressee identification for human-human-agent multiparty conversations in different proxemics*. In *Proc. of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, pp. 6:1–6:6. 2012.

[9] KATZENMAIER, M., R. STIEFELHAGEN, and T. SCHULTZ: *Identifying the addressee in human-human-robot interactions based on head pose and speech*. In *Proc. of the 6th ACM ICMI*, pp. 144–151. 2004.

[10] AKHTIAMOV, O., M. SIDOROV, A. KARPOV, and W. MINKER: *Speech and text analysis for multimodal addressee detection in human-human-computer interaction*. In *Proc. of the INTERSPEECH-2017*, pp. 2521–2525. 2017.

[11] RÖSNER, D., J. FROMMER, R. FRIESEN, M. HAASE, J. LANGE, and M. OTTO: *LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions*. In *Proc. of the 8th LREC*, pp. 96–103. Istanbul, Turkey, 2012.

[12] FROMMER, J., B. MICHAELIS, D. RÖSNER, A. WENDEMUTH, R. FRIESEN, M. HAASE, M. KUNZE, R. ANDRICH, J. LANGE, A. PANNING, and I. SIEGERT: *Towards emotion and affect detection in the multimodal last minute corpus*. In *Proc. of the 8th LREC*, pp. 3064–3069. Istanbul, Turkey, 2012.

[13] SIEGERT, I., D. PHILIPPOU-HÜBNER, K. HARTMANN, R. BÖCK, and A. WENDEMUTH: *Investigation of speaker group-dependent modelling for recognition of affective states from speech. Cognitive Computation*, 6(4), pp. 892–913, 2014.

[14] RÖSNER, D., M. KUNZE, M. OTTO, and J. FROMMER: *Linguistic analyses of the LAST MINUTE corpus*. In J. JANCSARY (ed.), *Proceedings of KONVENS 2012*, pp. 145–154. ÖGAI, 2012. Main track: oral presentations.

[15] BRUDER, C., C. CLEMENS, C. GLASER, and K. KARRER-GAUSS: *TA-EG – Fragebogen zur Erfassung von Technikaffinität*. Tech. Rep., FG Mensch-Maschine Systeme TU Berlin, 2009.

[16] PRYLIPKO, D., O. EGOROW, I. SIEGERT, and A. WENDEMUTH: *Application of Image Processing Methods to Filled Pauses Detection from Spontaneous Speech*. In *Proc. of the INTERSPEECH-2014*, p. s.p. Singapore, 2014.

[17] SIEGERT, I., M. HAASE, D. PRYLIPKO, and A. WENDEMUTH: *Discourse particles and user characteristics in naturalistic human-computer interaction*. In M. KUROSU (ed.), *Human-Computer Interaction. Advanced Interaction Modalities and Techniques*, vol. 8511 of *LNCS*, pp. 492–501. Springer, Berlin, Heidelberg, Germany, 2014.

[18] LANGE, J. and J. FROMMER: *Subjektives Erleben und intentionale Einstellung in Interviews zur Nutzer-Companion-Interaktion*. In *Proceedings der 41. GI-Jahrestagung*, vol. 192 of *Lecture Notes in Computer Science*, pp. 240–254. Bonner Köllen Verlag, Berlin, Germany, 2011.

[19] LEFTER, J., H. NEFS, C. JONKER, and L. ROTHKRANTZ: *Cross-corpus analysis for acoustic recognition of negative interactions*. In *Proc. of the 6th ACII*, pp. 132–138. Xian, China, 2015.

[20] EYBEN, F., M. WÖLLMER, and B. SCHULLER: *openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor*. In *Proc. of the ACM MM-2010*. 2010.

[21] TICKLE, A., S. RAGHU, and M. ELSHAW: *Emotional recognition from the speech signal for a virtual education agent*. *J. Phys.: Conf. Ser.*, 450, p. 012053, 2013.

[22] PFISTER, T. and P. ROBINSON: *Speech emotion classification and public speaking skill assessment*. In *Human Behavior Understanding*, vol. 6219 of *LNCS*, pp. 151–162. Springer, 2010.

[23] BÖCK, R., O. EGOROW, I. SIEGERT, and A. WENDEMUTH: *Comparative study on normalisation in emotion recognition from speech*. In P. HORAIN, C. ACHARD, and M. MALLEM (eds.), *Intelligent Human Computer Interaction: Proceedings of the 9th International Conference, IHCI 2017, Evry, France, December 11-13, 2017*, pp. 189–201. Springer International Publishing, Cham, 2017.

[24] HALL, M., E. FRANK, G. HOLMES, B. PFAHRINGER, P. REUTEMANN, and I. WITTEN: *The WEKA data mining software: An update*. *SIGKDD Explor. Newsl.*, 11(1), pp. 10–18, 2009.