# Speech Behavior Matters – Automatically Detect Device Directed Speech for the application of addressee-detection

Jun.-Prof. Ingo Siegert

Mobile Dialog Systems Group,
Institute for Information Technology and Communications
Otto von Guericke Universität Magdeburg

SPSC Webinar – 07.06.2021

# Outline

# Motivation

# Motivation

**Voice assistant systems recently receive increased attention**

# Motivation

### Voice assistant systems recently receive increased attention

- Microsoft Cortana had 133 million active users in 2016
- The echo Dot was the best-selling product on all of Amazon in the last three holiday seasons
- 72% of people owning a voice assistant often use them as part of their daily routine

# Motivation

## Voice assistant systems recently receive increased attention

- Microsoft Cortana had 133 million active users in 2016
- The echo Dot was the best-selling product on all of Amazon in the last three holiday seasons
- 72% of people owning a voice assistant often use them as part of their daily routine

## The ease of use is responsible for their attractiveness

By simply using speech commands users can:

- play music,
- search the web,
- create to-do and shopping lists,
- shop online,
- get instant weather reports, and
- control popular smart-home products.

# Motivation – Conversation Initiation

# Motivation – Conversation Initiation

**Nowadays:**

- Wake-up/Activation Word
- *Push-to-talk Button*

# Motivation – Conversation Initiation

## Nowadays:

- Wake-up/Activation Word
- *Push-to-talk Button*

## Problems of the wake-up word as the preferred method

- January 2017: **Alexa breakdown: Echo orders masses of doll's houses**
- September 2017: **Smart Home fraud: Neighbor is accepted to open the front door lock if a Siri Smart Lock**
- February 2018: **Amazon's Super Bowl Hack: Amazon has to "'mute"' its own wake-word in 3-6kHz frequencies**
- May 2018: **Embarrassing data breach – Alexa accidentally sends recorded conversation**

## Motivation – Conversation Initiation

### Nowadays:

- Wake-up/Activation Word
- *Push-to-talk Button*

### Problems of the wake-up word as the preferred method

- January 2017: **Alexa breakdown: Echo orders masses of doll's houses**
- September 2017: **Smart Home fraud: Neighbor is accepted to open the front door lock if a Siri Smart Lock**
- February 2018: **Amazon's Super Bowl Hack: Amazon has to "'mute'" its own wake-word in 3-6kHz frequencies**
- May 2018: **Embarrassing data breach – Alexa accidentally sends recorded conversation**

### An important aspect of interactions with voice assistants:

Detecting when the device should be activated
i.e. to distinguish human directed and device directed speech

# Research Questions

## Research Questions

1. How (good) do humans recognize the addressee?
2. How do recent automatic recognition systems perform?
3. What happens when we leave the lab?

# Utilized Datasets

# Voice Assistant Conversation Corpus (VACC) [Siegert et al., 2018]

# Voice Assistant Conversation Corpus (VACC) [Siegert et al., 2018]

## Voice Assistant Conversation Corpus (VACC)

- Based on interaction with a commercial voice assistant (ALEXA)
- User's self-reports on experiences during the interaction

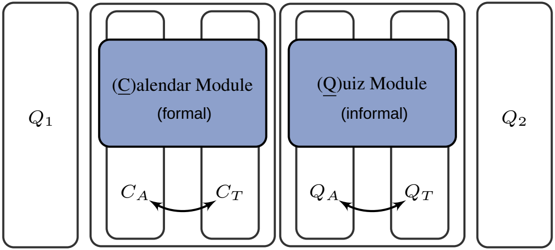
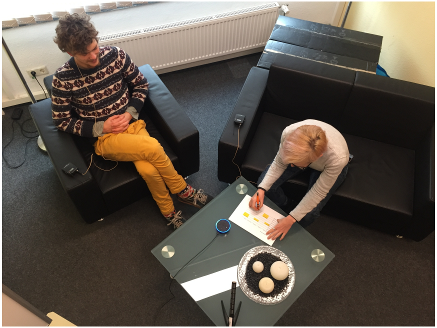



Figure: Sketch of the test procedure. $Q_1$ and $Q_2$ are the two rounds of the questionnaire. The order of the scenarios (calendar module and quiz module) is fixed. A and T denote the experimental conditions alone or together with a confederate.

# Recording setup



- Living-room like environment
- Natural communication atmosphere
- Amazon ALEXA Echo Dot (2. Generation)
- No video recording
    - 2x Neckband microphones
    - 1x Gun shot microphone
    - WAV uncompressed (44.1 kHz)

# VACC Dataset characteristics

| | |
|---|---|
| Participants | 27 |
| Sex | male 13 / female 14 |
| Age | Mean 24 (Std: 3.32) Min: 20; Max: 32 |
| Total duration | 17 h 07 min |

**Examples**

Example 1 Example 2 Example 3

# Differences in the complexity of HD/DD-dialogs

# Differences in the complexity of HD/DD-dialogs

Table: Example from SmartWeb Corpus

| HD | DD |
|---|---|
| Ey I'd love to go to Cologne to meet other fans. | Where does it go here in the city center? |
| He found over a hundred pubs. That's totally confusing. So where are we supposed to go to | Which country was the first Olympic champion in football? |

# Differences in the complexity of HD/DD-dialogs

Table: Example from SmartWeb Corpus

| HD | DD |
|---|---|
| Ey I'd love to go to Cologne to meet other fans. | Where does it go here in the city center? |
| He found over a hundred pubs. That's totally confusing. So where are we supposed to go to | Which country was the first Olympic champion in football? |

Table: Example from VACC

| HD | DD |
|---|---|
| yes best maybe you tell me when it's best for you | Alexa, do I have an appointment on Monday the 12. |
| I've always done it this way and asked when he was born and when he died | when was Martin Luther King born |

# Differences in the complexity of HD/DD-dialogs

Table: Example from SmartWeb Corpus

| HD | DD |
|---|---|
| Ey I'd love to go to Cologne to meet other fans. | Where does it go here in the city center? |
| He found over a hundred pubs. That's totally confusing. So where are we supposed to go to | Which country was the first Olympic champion in football? |

Table: Example from VACC

| HD | DD |
|---|---|
| yes best maybe you tell me when it's best for you | Alexa, do I have an appointment on Monday the 12. |
| I've always done it this way and asked when he was born and when he died | when was Martin Luther King born |

**Mismatch of the dialog complexity could influence the recognition problem!**

Whether the counterpart is a human or a technical system

# Restaurant Booking Corpus [Siegert et al. 2019]

# Restaurant Booking Corpus [Siegert et al. 2019]

## Design

- Interactions with technical systems or humans via simulated telephone
- Explicit design of DD-/HD-dialogs with
  - Same type of conversation
  - Same vocabulary
  - Same content

# Restaurant Booking Corpus [Siegert et al. 2019]

## Design

- Interactions with technical systems or humans via simulated telephone
- Explicit design of DD-/HD-dialogs with
  - Same type of conversation
  - Same vocabulary
  - Same content

## Task

- Reservations in 3 restaurants
- Various constraints
- Interlocutor is either
  - a simple technical system
  - an advanced technical system
  - a human Interlocutor

# Restaurant Booking Corpus [Siegert et al. 2019]

## Design

- Interactions with technical systems or humans via simulated telephone
- Explicit design of DD-/HD-dialogs with
  - Same type of conversation
  - Same vocabulary
  - Same content

## Task

- Reservations in 3 restaurants
- Various constraints
- Interlocutor is either
  - a simple technical system
  - an advanced technical system
  - a human Interlocutor

## Data set

- 30 participants (10 m/ 20f)
- 5 h 37 min
- 4835 utterances

| | |
|---|---|
| TS1 | 797 |
| TS2 | 637 |
| H | 789 |

Example 1 Example 2 Example 3

# Research Questions

# Research Questions

1. How (good) do humans recognize the addressee?
2. How do recent automatic recognition systems perform?
3. What happens when we leave the lab?

# How (good) do humans recognize the addressee?

# How (good) do humans recognize the addressee?

**[Katzenmaier et al., 2004, Jovanovic et al., 2006, Beyan et al., 2016]**

The majority of studies refer to visual (gaze) or lexical (wake-up word) cues.

# Research methods

# Research methods

## Subjective analysis (VACC)

- Open and closed questions
  - Experiencing changes in speaking style
  - Verbalisation of differences
- Summarizing qualitative content analysis [Mayring, 2014]

# Research methods

## Subjective analysis (VACC)

- Open and closed questions
  - Experiencing changes in speaking style
  - Verbalisation of differences
- Summarizing qualitative content analysis [Mayring, 2014]

## Objective Analysis (VACC,RBC)

- Human Annotation
  - GER and NON-GER
  - 10 annotators each
  - random pre-selection
  - without lexical cues
- Calculating recall measure

# Results – Subjective analysis

# Results – Subjective analysis

## Interaction with human partner

- „frei und unbekümmert“ (B), „intuitiv“ (X)
- „gesprochen wie immer“ (G), „keine großen Gedanken gemacht“ (M), weil Kommunikation mit Menschen geläufig ist
- „persönlich[eres]“, „freundlich[eres]“ Sprechen (E)

# Results – Subjective analysis

## Interaction with human partner

- „frei und unbekümmert“ (B), „intuitiv“ (X)
- „gesprochen wie immer“ (G), „keine großen Gedanken gemacht“ (M), weil Kommunikation mit Menschen geläufig ist
- „persönlich[eres]“, „freundlich[eres]“ Sprechen (E)

## Interaction with Alexa

- kaum „intuitiv“ (AB) erlebt,
- „schwieriger zu kommunizieren“ (P), „unfrei“ (B), „kein Dialog“ (J)
- „[Betonung und Lautstärke] eher anders als ich es mit jemanden in der realen Welt gemacht hätte“ (M)

# Results – Objective analysis

# Results – Objective analysis

| UAR | GER | NON-GER |
|---|---|---|
| VACC | 82.27% | 71.68% |
| | | |
| | | |

# Results – Objective analysis

| UAR | GER | NON-GER |
|---|---|---|
| VACC | 82.27% | 71.68% |
| SVC | 85.45% | 72.43% |

# Results – Objective analysis

| UAR | GER | NON-GER |
|---|---|---|
| VACC | 82.27% | 71.68% |
| SVC | 85.45% | 72.43% |
| **RBC** | **60.54%** | **53.57%** |

# How do recent automatic recognition systems perform?

# How do recent automatic recognition systems perform?

| [Baba et al., 2012] | [Shriberg et al., 2012] | [Tsai et al., 2015] | [Batliner et al., 2008] |
|---|---|---|---|
| 2 Persons | 2 Persons | 2-3 Persons | 2 Persons |
| Animated character | "Conversational Browser" | Computer | Computer |
| Decision-making | Formal interaction | Quiz | Information retrieval |
| 6 features ($F_0$, intensity, speech rate) | Energy(-contour), speech rate | 47 features (energy-contour) | duration, energy, F0, length of pauses |
| SVM | GMM | Adaboost | LDA |
| 80.7% (Accuracy) | 12,63% (EER) | 13.88% (EER) | 74.2% (UAR) |

# Own Recognition Experiments

# Own Recognition Experiments

**Baseline** [Siegert et. al, 2018,2019] **VACC/RBC**

- linear SVM (emobase features)

# Own Recognition Experiments

## Baseline [Siegert et. al, 2018,2019] VACC/RBC

- linear SVM (emobase features)

## Metaclassifier [Akhtiamov et. al, 2019,2020] VACC/RBC

- linear SVM from ComParE (LDDS + func)
- radial SVM from ASR configuration
- LSTMs from LLDs
- LSTMs from raw-audio

# Own Recognition Experiments

### Baseline [Siegert et. al, 2018,2019] VACC/RBC

- linear SVM (emobase features)

### Metaclassifier [Akhtiamov et. al, 2019,2020] VACC/RBC

- linear SVM from ComParE (LDDS + func)
- radial SVM from ASR configuration
- LSTMs from LLDs
- LSTMs from raw-audio

### RNNs with attention layer [Baumann & Siegert, 2020] RBC

- features: MFCCs and FFV + phone identities
- RNNs for each segment, attention layer over different segment sizes

# Own Recognition Experiments

## Baseline [Siegert et. al, 2018,2019] VACC/RBC

- linear SVM (emobase features)

## Metaclassifier [Akhtiamov et. al, 2019,2020] VACC/RBC

- linear SVM from ComParE (LDDS + func)
- radial SVM from ASR configuration
- LSTMs from LLDs
- LSTMs from raw-audio

## RNNs with attention layer [Baumann & Siegert, 2020] RBC

- features: MFCCs and FFV + phone identities
- RNNs for each segment, attention layer over different segment sizes

## Continuous Learning Framework [Siegert et. al (submitted CSR)] RBC

- speaker-dependend architecture, re-train on few samples

# Automatic Recognition Results

# Automatic Recognition Results

| | VACC | | RBC | |
|---|---|---|---|---|
| | UAR | abs. $\Delta$ | UAR | abs. $\Delta$ |
| Human Annotation (NON-GER) | 71.68% | – | 53.57% | – |
| Baseline (linear SVM) | | | | |
| Metaclassifier | | | | |
| RNNs + attention layer | | | | |
| Continuous Learning Framework | | | | |

# Automatic Recognition Results

| | VACC | | RBC | |
|---|---|---|---|---|
| | UAR | abs. $\Delta$ | UAR | abs. $\Delta$ |
| Human Annotation (NON-GER) | 71.68% | – | 53.57% | – |
| Baseline (linear SVM) | 85.38% | 13.70% | 52.02% | -1.55% |
| Metaclassifier | | | | |
| RNNs + attention layer | | | | |
| Continuous Learning Framework | | | | |

# Automatic Recognition Results

| | VACC | | RBC | |
|---|---|---|---|---|
| | UAR | abs. $\Delta$ | UAR | abs. $\Delta$ |
| Human Annotation (NON-GER) | 71.68% | – | 53.57% | – |
| Baseline (linear SVM) | 85.38% | 13.70% | 52.02% | -1.55% |
| Metaclassifier | 89.10% | 17.42% | 62.70% | 9.13% |
| RNNs + attention layer | | | | |
| Continuous Learning Framework | | | | |

# Automatic Recognition Results

| | VACC | | RBC | |
|---|---|---|---|---|
| | UAR | abs. $\Delta$ | UAR | abs. $\Delta$ |
| Human Annotation (NON-GER) | 71.68% | – | 53.57% | – |
| Baseline (linear SVM) | 85.38% | 13.70% | 52.02% | -1.55% |
| Metaclassifier | 89.10% | 17.42% | 62.70% | 9.13% |
| RNNs + attention layer | – | – | 65.50% | 11.93% |
| Continuous Learning Framework | | | | |

# Automatic Recognition Results

| | VACC | | RBC | |
|---|---|---|---|---|
| | UAR | abs. $\Delta$ | UAR | abs. $\Delta$ |
| Human Annotation (NON-GER) | 71.68% | – | 53.57% | – |
| Baseline (linear SVM) | 85.38% | 13.70% | 52.02% | -1.55% |
| Metaclassifier | 89.10% | 17.42% | 62.70% | 9.13% |
| RNNs + attention layer | – | – | 65.50% | 11.93% |
| Continuous Learning Framework | – | – | 85.77% | 32.20% |

# Summary & Limitations

# Summary & Limitations

## Summary

1. How (good) do humans recognize the addressee? ✓/✗

## Summary & Limitations

### Summary

1. How (good) do humans recognize the addressee? ✓/✗
2. Automatic recognition performance? ✗, speaker-dependancy ✓

# Summary & Limitations

## Summary

1. How (good) do humans recognize the addressee? ✓/✗
2. Automatic recognition performance? ✗, speaker-dependancy ✓
3. Prosodic information help to distinguish HD and DD utterances ✓

# Summary & Limitations

## Summary

1. How (good) do humans recognize the addressee? ✓/✗
2. Automatic recognition performance? ✗, speaker-dependancy ✓
3. Prosodic information help to distinguish HD and DD utterances ✓

## Limitations

- Number of participants
- No video recording
- Dialog complexity/Length of interaction
- Non parallel interaction of human and system
- Lab environment

# What happens when we leave the lab?

# What happens when we leave the lab?

## Limitations of Lab environment

- Participants try to act as good participants
- Hard to get participants' real feelings
- Participants need a task

## Need for unrestricted data in public environment

- people talk voluntary,
- people talk unrestricted,
- people talk without fear of being observed/recorded, and
- people themselves determine beginning and end of the conversation.

## “Alexa in the wild” – Voice Assistant Conversations in the wild (VACW) Corpus [Siegert, 2020]

# “Alexa in the wild” – Voice Assistant Conversations in the wild (VACW) Corpus [Siegert, 2020]

## “MS Wissenschaft”

- May until October 2019
- 31 cities in Germany and Austria
- Stay of 3 to 5 days for each city
- Exhibition is aimed at school classes but also at interested adults
- More than 85 000 people with more than 500 classes visited exhibition

# “Alexa in the wild” – Voice Assistant Conversations in the wild (VACW) Corpus [Siegert, 2020]

## “MS Wissenschaft”

- May until October 2019
- 31 cities in Germany and Austria
- Stay of 3 to 5 days for each city
- Exhibition is aimed at school classes but also at interested adults
- More than 85 000 people with more than 500 classes visited exhibition

# Dataset overview

# Dataset overview

| | |
|---|---|
| Duration | 39.9h |
| # Visitor utterances | 32 758 |
| # Sessions | 7 144 |
| Language | German |
| Annotation | transcriptions, topics |

Table: Key characteristics of the VACW dataset.

**Examples**

Example 1 Example 2 Example 3

# First Analyses I

# First Analyses I

| Topics | Frequency |
|---|---|
| Quiz-Questions | 41.3% |
| Other-Questions | 10.1% |
| Alexa features | 16.0% |
| Time/Date | 7.4% |
| Music | 5.6% |
| Playing around | 3.2% |
| Weather | 1.4% |
| Inappropriate | 1.4% |
| Saluations | 0.8% |
| Games | 0.4% |
| Movie/TV | 0.2% |
| Recommendations | 0.1% |
| Other | 12.1% |

Table: Types of visitor interactions with Alexa during the exhibition, with examples and frequency.

| Activation word | Occurrences |
|---|---|
| Alexa | 8 732 |
| Alexa (multiple times) | 314 |
| Hey Alexa | 16 |
| Hi Alexa | 1 |
| Hey Siri | 3 |
| Hey Google | 3 |
| Google | 1 |

Table: Distributions of different ”activation“ words. As Alexa sometimes allow to utter follow-up requests, not all utterances need an activation word and therefore this number is smaller than the total number of utterances.

# First Analyses II

# First Analyses II

## Other remarkable observations

- group interactions
- asking regarding surveillance
- asking for Alexa to be his/her friend and for marriage
- giving good bye messages to Alexa
- Swearwords and other non appropriate words
- 15% of user request could not been solved (”I did not understand”)

# More to come...

# Conclusion & Outlook

# Conclusion & Outlook

## Conclusion

- Actual wake-word activation sometimes fails
- Humans use additional cues (gaze, prosody) to code the addressee
- Addressee-detection should use this information
- Even with challenging data a prosody-only AD possible
- If individual differences are taken into account

# Conclusion & Outlook

## Conclusion

- Actual wake-word activation sometimes fails
- Humans use additional cues (gaze, prosody) to code the addressee
- Addressee-detection should use this information
- Even with challenging data a prosody-only AD possible
- If individual differences are taken into account

## Outlook

- Analyse individual addressee behavior
- Analyse single factors of addressee behavior
  - Appearance
  - System's Voice
  - User type
  - ...
- Larger datasets needed
- Testing under real conditions (imperfect audio, compression, etc.)

## Thank you for your attention

### Research Collaborators

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

EIT FAKULTÄT FÜR ELEKTROTECHNIK UND INFORMATIONSTECHNIK

Mobile Dialog Systeme

Oleg Akhtiamov, Timo Baumann, Ralph Heinemann,
Julia Krüger, Jannik Nietzold, Eran Raveh,
Norman Weißkirchen, Andreas Wendemuth

# Literature


Akhtiamov, Oleg, Ingo Siegert, et al. (Sept. 2019). “Cross-Corpus Data Augmentation for Acoustic Addressee Detection”. In: Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue. Stockholm, Sweden: Association for Computational Linguistics, pp. 274–283. DOI: 10.18653/v1/W19-5933.

— (2020). “Using Complexity-Identical Human- and Machine-Directed Utterances to Investigate Addressee Detection for Spoken Dialogue Systems”. In: Sensors 20.9, p. 2740. DOI: https://doi.org/10.3390/s20092740.

Baumann, Timo and Ingo Siegert (2020). “Prosodic addressee-detection: ensuring privacy in always-on spoken dialog systems”. In: Mensch und Computer 2020 - Tagungsband. Ed. by Florian Alt, Stefan Schneegass, and Eva Hornecker. New York: ACM, pp. 195–198. DOI: 10.1145/3404983.3410021.

Siegert, Ingo (May 2020). ““Alexa in the wild” – Collecting Unconstrained Conversations with a Modern Voice Assistant in a Public Environment”. In: Proceedings of The 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, pp. 608–612.

Siegert, Ingo and Julia Krüger (2018). “How do we speak with ALEXA - subjective and objective assessments of changes in speaking style between HC and HH conversations”. In: Kognitive Systeme (1).

Siegert, Ingo, Julia Krüger, et al. (May 2018). “Voice Assistant Conversation Corpus (VACC): A Multi-Scenario Dataset for Addressee Detection in Human-Computer-Interaction using Amazon’s ALEXA”. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Ed. by Hanae Koiso and Patrizia Paggio. Miyazaki, Japan: European Language Resources Association (ELRA).

Siegert, Ingo, Jannik Nietzold, et al. (2019). “The Restaurant Booking Corpus – content-identical comparative human-human and human-computer simulated telephone conversations”. In: Elektronische Sprachsignalverarbeitung 2019. Tagungsband der 30. Konferenz. Vol. 93. Studientexte zur Sprachkommunikation. Dresden, Germany: TUDpress, pp. 126–133.