
FAKULTÄT FÜR
ELEKTROTECHNIK UND
INFORMATIONSTECHNIK

# Data protection and pseudonymisation of speech data

2nd LDS Technology Workshop,
January 29th, 2024

**Ingo Siegert**
**Mobile Dialog Systems**

# Processing of (Speech) Data under the GDPR

- Personal data → any information related to an identified or identifiable natural person.

- Anonymised data
  - the data subject cannot be identified by any means reasonably likely to be used.
  - no longer personal data
    BUT: high cost and loss of valuable information

- **Alternative:** Pseudonymisation
  - data can no longer be attributed to a specific person without the use of additional information
  - the key shall be kept separately from the data,
    - subject to technical and organisational measures to prevent re-identification (consider a Data Breach Policy)
    - a safeguard for the rights and freedoms of data subjects

# Pseudonymisation and Transparency

- data subject should be provided with information
  - about the processing,
  - also if date is not obtained directly from user
- **Exception**: provision of the information would require disproportionate effort.

- Assessing the effort:
  - 1) the number of data subjects;
  - 2) the age of the data;
  - 3) adopted safeguards (e.g. pseudonymisation)

# Takeaway Message on Pseudonymisation

- Pseudonymisation ≠ Anonymisation

- Pseudonymised data are still personal data,
  - but they are much easier to process, especially for research purposes:
  - even if they were originally collected for a different purpose
    - they can be re-used for other research projects
    - they can be stored for extended periods
    - rights of data subjects wrt. to the data are limited
    - the obligation to inform data subjects about the processing can be avoided.

# Pseudonymisation of Speech Data: A Case Study

- Recorded in cooperation with a German call centre
  - Real telephone-based conversations
  - Four agents recorded on a daily basis
- Using recording carrel,
  - Minimize surrounding noise
- Audio stream of both, agent and caller
  - If caller gives consent

Final Dataset contains:

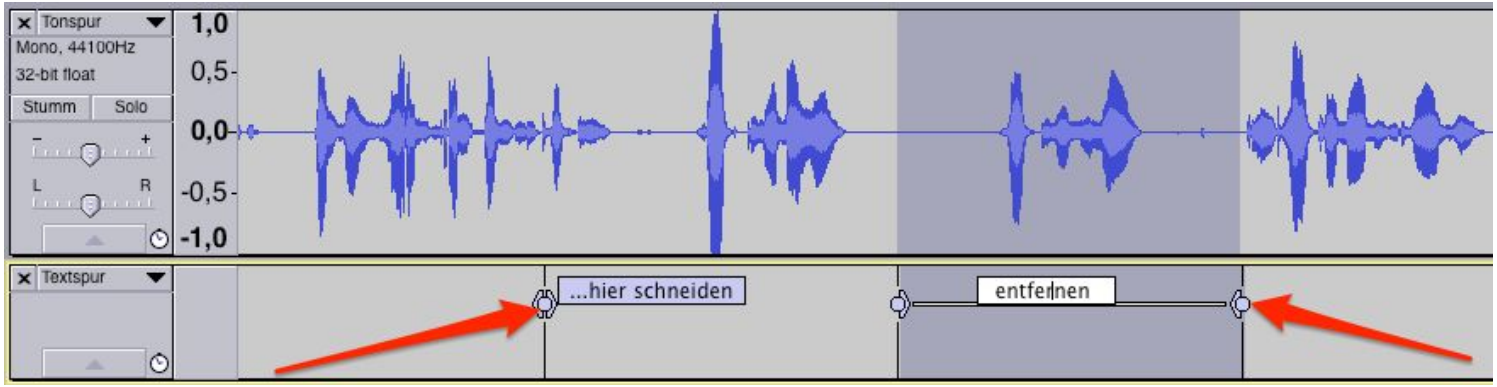
- 93 hours of speech data
- Pseudonymisation and segmentation in user turns needed!

# Removal of directly identifying information of the caller

- Done in parallel to the dialog/turn segmentation
- Ease the listening effort
- Call-Center in-house listeners employed (NDA)
- Had to listen to all recordings
  - Replace identifying information by silence
    - Name, Contract-Number, Bank-Details, Fees, etc.

Used Audacity

- easy interface, key-codes
- no storage of further information
- Researcher only get the cleaned acoustics with random ID



# Pseudonymisation – UserID and Cross-Studies

- In bigger research centres, participant pools are used
  - Compare same participants across different experiments?
  - Is it desirable? Of course!
    - longitudinal studies, repeated behaviour, linking of different measurements
    - Remember: data still handled with care and participants need to be informed that the data is linked to previous experiments (in the same organisation)
  - How to identify identical participants?
    - Lazy method: always use same way to generate UserID:
      - not error prone, could be doubled, directly reveal personal information
    - Proposed method: use hashing algorithm (SHA-2)
      - build from personal information, e.g. IngoSMay1983 → 10e9597f9c1712bb7d2a32e694079f28b661b39ab349c1db75db5c25c71e8053
      - not reversible!!!

LinkedIn

Homepage

# Thanks for your attention

Jun.Prof. Dr. Ingo Siegert
Mobile Dialog Systems Group
Otto von Guericke University Magdeburg

https://www.ingo-siegert.de/
ingo.siegert@ovgu.de

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG