The Legal and Ethical issues Workshop @LREC2022, pages 17–21
Marseille, 20 June 2022

Pseudonymisation of Speech Data as an Alternative Approach to GDPR Compliance

Paweł Kamocki[1] **and Ingo Siegert**[2]
[1]Leibniz Institut für Deutsche Sprache, Mannheim, Germany
[2]Mobile Dialog Systems, Otto von Guericke University Magdeburg, Germany
kamocki@ids-mannheim.de, siegert@ovgu.de

Abstract

The debate on the use of personal data in language resources usually focuses — and rightfully so — on anonymisation. However, this very same debate usually ends quickly with the conclusion that proper anonymisation would necessarily cause loss of linguistically valuable information. This paper discusses an alternative approach — pseudonymisation. While pseudonymisation does not solve all the problems (inasmuch as pseudonymised data are still to be regarded as personal data and therefore their processing should still comply with the GDPR principles), it does provide a significant relief, especially — but not only — for those who process personal data for research purposes. This paper describes pseudonymisation as a measure to safeguard rights and interests of data subjects under the GDPR (with a special focus on the right to be informed). It also provides a concrete example of pseudonymisation carried out within a research project at the Institute of Information Technology and Communications of the Otto von Guericke University Magdeburg.

Keywords: Pseudonymisation, GDPR, Personal Data, Speech Data

1. Introduction

In European law, personal data are defined in a very broad manner as 'any information related to an identified or identifiable natural person'. This definition, currently in §4 of the GDPR, is in fact much older than the GDPR itself, and can be traced back to the 1981 Council of Europe's Convention 108, or even to the 1977 German Federal Data Protection Act (itself inspired by the 1970 Data Protection Act of the State of Hessen). This very general and broad approach is the cornerstone of European privacy law.

Under this approach, even information that is not nominative (i.e. does not contain the person's name and surname) or directly identifying (e.g. a social security number) should be regarded as personal data, as long as it can be related to a person. Therefore, a huge part of language data, especially in speech and multi-modal resources, fall within the scope of data protection laws. As such, the processing of such data should abide by the GDPR principles of lawfulness, fairness, transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity, confidentiality, and accountability. A good overview on reflections on legal and technical issues regarding speech data and GDPR is given in (Nautsch et al., 2019).

These principles no longer apply to data that have been anonymised, i.e. processed in such a manner that the person they originally referred to can no longer be identified 'by any means likely reasonably to be used'. However, anonymisation should be permanent and irreversible (WP29 (Article 29 Data Protection Working Party), 2014), which almost always entails a loss of potentially valuable linguistic information (Siegert et al., 2020). Moreover, taking into account the growing availability of online data that can be used to re-identify the person, the technical standard for anonymisation (set high by the 2014 WP29 opinion on anonymisation techniques) is constantly getting higher. Therefore, apart from being a technological and organisational challenge (with many tasks that still have to be performed manually), anonymisation is necessarily a costly procedure.

Pseudonymisation, which should be clearly distinguished from anonymisation, may be an alternative solution. Rather than permanently breaking the relation between the person and the data, pseudonymisation consists of the processing of the data in such a manner that it can no longer be attributed to a specific person without the use of additional information (e.g. a pseudonym or an ID number). This additional information (which can be referred to as 'the key') shall be kept separately from the data, and be subject to technical and organisational measures to prevent re-identification of data subjects (cf. definition of pseudonymisation in §4 of the GDPR).

Under the GDPR, pseudonymisation is one of the possible safeguards for the rights and freedoms of data subjects (Section 2), which, if applied correctly, reduces the legal burden at various stages of data processing (also, for example, regarding the data subjects' right to information; Section 3). It is therefore an interesting option to consider in research projects, for example in the field of speech data (Section 4).

2. Pseudonymisation in the GDPR

Unlike the 1995 Personal Data Directive (in force until 2018), the GDPR explicitly introduces pseudonymisation as a safeguard that can 'reduce the risks to the data subjects concerned and help controllers and processors to meet their data-protection obligations' (Recital 28 of the GDPR). This has several practical consequences,

especially regarding the so-called 'purpose extension' (WP29 (Article 29 Data Protection Working Party), 2013), and the processing of personal data for research purposes. Purpose extension is the principle according to which data lawfully collected for one purpose can be subsequently re-used (without e.g. the need to obtain new consent from data subjects) for a 'compatible purpose'. By means of exception, scientific research shall always be regarded as a compatible purpose (as per Article 5.1 (b) of the GDPR). However, if the purpose is different from scientific research, then it is for the data controller to assess the compatibility of the new purpose with the initial purpose. Article 6.4 of the GDPR lists five elements that can be taken into account in this assessment (the list is not exhaustive); the existence of safeguards such as pseudonymisation is one of them. Therefore, pseudonymisation facilitates the use of lawfully collected data for a new purpose, as it enlarges the scope of 'compatible' purposes.

When the processing is carried out for research purposes, Article 89 of the GDPR allows the Member States to adopt a number of exceptions and derogations from the general data protection framework. These derogations concern e.g. the purpose limitation principle (scientific research is always regarded as a 'compatible purpose'), the storage limitation (for research purposes, data can be stored for longer than 'necessary'), as well as some rights of data subjects (information, erasure, right to object). An important caveat, however, is that in order to be able to qualify for all these derogations, the processing should be not only carried out exclusively for scientific research purposes (including commercial research), but also it should be subject to 'appropriate safeguards'. Article 89 of the GDPR expressly lists pseudonymisation as an example (the only example) of such a safeguard. Arguably, pseudonymisation is in most cases the cheapest safeguard, and the easiest to implement.

Before we discuss a concrete example of pseudonymisation, it should be pointed out that pseudonymisation, in order to meet the requirements of the GDPR, should involve appropriate technical and organisational security measures to prevent unauthorised access to the 'key' and identification of data subjects. Such organisational security measures, as per Articles 32 and following of the GDPR, can include a Data Breach Policy — an internal procedure to follow in case of an event which may constitute a data breach, and the criteria to determine the related risks for data subjects. It should be reminded here that a breach, if it is likely to result in a risk for the rights and freedoms of natural persons, should be notified to the supervisory authority, and if the risk is high — also communicated to data subjects.

3. Data Subject's Right to Information under the GDPR

As discussed in the previous section, pseudonymised data are still to be regarded as personal data, and therefore their processing should in principle still observe the General Data Protection Regulation. This means that, among other obligations, data subjects can still exercise their rights, unless a statutory exception applies.

Information is the most fundamental right of data subjects. According to Article 12 of the GDPR, the controller shall take appropriate measures to inform data subjects about the processing in 'a concise, transparent, intelligible and easily accessible form, using clear and plain language, in particular for any information addressed specifically to a child'. The information should be provided in writing, including in electronic form. Oral transmission of information is not excluded, but it is much harder to document, which is especially important since according to the accountability principle, the controller should be able to demonstrate compliance with the GDPR. Moreover, the sheer amount of information that, according to Articles 13 and 14, should be communicated to the data subject (see below), transmission in writing is also more practicable.

Importantly, data subjects shall be provided with information regardless of whether their consent is asked for in the process and regardless of whether the data were obtained directly from them or from other sources (including publicly available sources such as public LinkedIn profiles). In the first case (data obtained directly from the subject), the information should be provided at the time when the data is obtained; in the second (data obtained from other sources) - within a reasonable period of time, but no later than a month after the data have been obtained, or - if the data are disclosed to another recipient (e.g. shared with another research team) - at the latest at the moment of this disclosure.

Regarding the elements that data subjects should be provided with, the GDPR contains two lists: Article 13 applies when the data are collected directly from the data subject; Article 14 - in other cases. For the most part, both lists overlap; they both include such elements as (among others) the identity and contact details of the data controller, the purposes and the legal basis for the processing (including, where this basis applies, the legitimate interest pursued by the controller), the period for which the data will be stored in unanonymised form (or at least how the period will be determined), the persons (or categories of persons) the data will be disclosed to (recipients) and, if applicable, intended transfers of the data outside the European Economic Area. Both Article 13 and 14 also require information about the rights of the data subject, including the right to withdraw consent (if the processing is based on consent) or to lodge a complaint with a supervisory authority. The most important difference in the content of information between the two articles is that where the data are not obtained directly from the data subject (Article 14), he or she has to be informed about the categories of data collected and about the source it was obtained from (including information on whether the source is publicly available).

It shall be noted that in practice, most of these elements

Table 1: Overview of selected information provided to data subjects.

	Data collected directly from subject (13 GDPR)	Data obtained from another source (14 GDPR)
When to inform	At the time of collection	Max. 1 month after obtaining data
Exception 1	Subject already has the information	
Exception 2		provision is impossible or requires disproportionate effort
Controller's identity and contact	+	
Data protection officer's contact	+	
Purpose(s) of the processing	+	
Categories of processed data	-	+
Legal basis of the processing (or legitimate interest)	+	
Recipients	+	
Transfers outside European Economic Area, if intended	+	
Data retention period (or criteria to determine it)	+	
Right to lodge a complaint	+	
Right to withdraw consent	+	
Whether the provision of data is required (by law or by contract), and consequences of refusal	+	-
Source data was obtained from	-	+
Existence of automated decision-making (see 22 GDPR)	+	

can be covered in a boilerplate text (with some modifications to fit specific scenarios), it is therefore highly recommendable to work on a re-usable model for an information form (sometimes referred to as 'consent form', rather mistakenly, since the information has to be provided even when there is no need to obtain consent, i.e. when processing is based on other grounds, such as legitimate interests).

The main interest in distinguishing between the situation when the data are obtained from the data subject and when they are obtained from other sources is in the exceptions. In the first scenario, Article 13.4 allows for only one exception: the information does not have to be provided when the data subject already has it. However, when the data are not obtained directly from the data subject, there is considerably more leeway; the obligation to provide information can be derogated from (Article 14.5) also when it proves impossible or would involve a disproportionate effort or in so far as the provision of information is likely to render impossible or seriously impair the achievement of the objectives of that processing. This is particularly relevant when the processing (of the data) is carried out for research purposes, and the application. In assessing whether the obligation can be derogated based on disproportionate efforts, account should be taken of three elements (WP29 (Article 29 Data Protection Working Party), 2018): the number of data subjects (the higher the number, the bigger the effort), the age of the data (the older the data, the bigger the effort) and any appropriate safeguards adopted. In this approach, the use of safeguards such as pseudonymisation may be a factor that 'tilts the scales' on the side of the derogation. The differences between Article 13 and Article 14 are summarized in Table 1.

However, even if the derogation from the obligation to provide information applies, transparency of the processing should still be observed. In such case, the controller should take appropriate measures to protect the data subject's rights and freedom, e.g. by making the information about the processing publicly available. In the context of research projects, when the data are collected directly from the subjects, and where measures such as pseudonymisation are applied, publishing a note with all required elements on the institution's (or the project's) website would often be enough to comply with the obligation.

4. Pseudonymisation of Speech Data: A Case Study

Naturalistic data recordings are an important resource for speech-based analyses. Therefore, data should be of high quality, including long and elaborate interactions, non-verbal events, and having a reliable and versatile emotion annotation. Ideally, the data set should contain contextual information about the speakers, such as age, sex, or personality traits, see (Böck et al., 2019).

The reported case study concerns a dataset recorded under a transfer project within the DFG-funded SFB/TRR-62 "A Companion Technology for cognitive technical systems"[1] at the Institute of Information Technology and Communications of the Otto von Guericke University Magdeburg in collaboration with a German call centre agency. The aim was to automatically support the agent in the handling of affective customer signals. It was aimed to give feedback to the agents regarding their dialogue with the customer and to give suggestions for customer-oriented dialogues. As call centre agents are mostly dealing with the factual level of the conversations and are rather insensitive to signals on the relational level (Watzlawick et al., 1967). The project ran from 2015 until 2016.

To support this hypothesis and develop a suitable recognition system, suitable data of sufficient amounts have to be available. To exclude side effects, which prevent a satisfactory classification performance on the expected less expressive emotional expressions, data having the same context and the same acoustic conditions are necessary (Douglas-Cowie et al., 2005; Zeng et al., 2009). Therefore, a larger data collection to train the recognition models and to obtain a sufficient number of different caller and agent behaviour was conducted at the beginning of the project. This recording has on the one hand to protect the personal data of both the agent and the caller and on the other hand allowing to record and analyse the recorded voice data.

The audio stream of both agent and caller was recorded. To later inspect the recordings for peculiarities, the agent was video-recorded as well. The callers were informed about the fact that the call was being recorded by preceding information that "the conversation is recorded due to quality reasons and the customers can refuse to accept this recording at any time". The agents took part voluntarily and their name has never been disclosed to the academic partner. As it is known that the emotional reaction is heavily dependent on personality ((Larsen and Ketelaar, 1991)), the agent's evaluation regarding the Big Five personality traits ((Costa and McCrae, 1995)) and the stress-coping questionnaire ((Jahnke et al., 2002)) are stored as well. As for the agents, also age, gender, and personality information were recorded, an agent code (Agent1 ... Agent4) was used to pseudonymise this information.

To conduct the recording, a separate recording carrel was established. Thereby surrounding noise could be minimized, for the video recordings the privacy of people not involved could be preserved and a uniformly illuminated scene was enabled. All of these recordings took place in-house at the call centre agency. Furthermore, a special button was installed to interrupt the recording if a customer withdraws the initially given consent for recording.

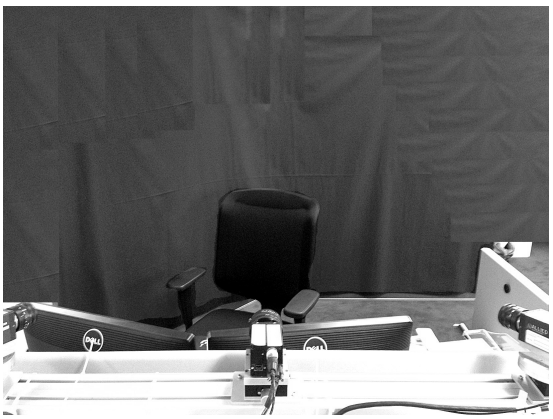

Figure 1: Picture of the separate recording carrel

The dataset ((Siegert and Ohnemus, 2015)) comprised real telephone-based conversations of in total 1 447 dialogues with 46 610 turns, which comprises approx. 93 hours of speech data. The topics of the calls range from simple informative calls and notifications of changes of customer data to complaint calls. In order to enable a comprehensive analysis of the material, four agents were selected, and their conversations were recorded on a daily basis.

As the phone calls were authentic customer dialogues, they had to be "pre-anonymised" first. Therefore, specially trained employees carefully listened to all recordings. All passages where personal information was disclosed were replaced by corresponding silence passages. The employees used Audacity for this task. Although most of the procedure could be sped up by using specialized keyboard shortcuts, this task had a processing time from 6 times the original recording time. To pseudonymise the remaining data, each recorded dialogue is stored under a consecutive number. A separate file holds the detailed information of the specific recording time for each dialogue. This file connects the consecutive number of each dialogue (the filename, e.g. 0001.wav) with its recording time (e.g. 31. February 2016, Dialogue 55). This file is stored on a separate external hard disk, in a locked cabinet, where only the lead scientists have access.

5. Conclusion

Pseudonymisation should not be mistaken for anonymisation; pseudonymised data are still to be considered personal data, but if the pseudonymisation is done correctly (also with regard to organisational and technical security measures to prevent de-identification), it may

[1] http://www.sfb-trr-62.de/

allow for the data to be lawfully processed for scientific research purposes, without losing all the relevant information. It may also be less costly than anonymisation. The pseudonymised data allows for research on prosodic-acoustic analyses by distributing extracted characteristics for acoustic modelling and by allowing in-house listener evaluations. Pseudonymisation of audio data is still an open issue, especially, as techniques to anonymize the speaker (obfuscating the speaker ID) while preserving relevant speech and emotional content is still under development (Sinha and Siegert, 2022; Tomashenko et al., 2021). Therefore, it should always be considered as an alternative way to GDPR compliance for scientific research projects, especially those involving processing of speech data, which are particularly hard to anonymise.

6. Bibliographical References

Böck, R., Egorow, O., Höbel-Müller, J., Requardt, A. F., Siegert, I., and Wendemuth, A., (2019). *Anticipating the User: Acoustic Disposition Recognition in Intelligent Interactions*, pages 203–233. Springer International Publishing, Cham.

Costa, P. T. and McCrae, R. R. (1995). Domains and Facets: Hierarchical Personality Assessment Using the Revised NEO Personality Inventory. *J Pers Assess*, 64:21–50.

Douglas-Cowie, E., Devillers, L., Martin, J.-C., Cowie, R., Savvidou, S., Abrilian, S., and Cox, C. (2005). Multimodal databases of everyday emotion: facing up to complexity. In *Proc. of the INTERSPEECH-2005*, pages 813–816, Lisbon, Portugal.

Jahnke, W., Erdmann, G., and Kallus, K. (2002). *Stressverarbeitungsfragebogen mit SVF 120 und SVF 78*. Hogrefe, Göttingen, Germany, 3 edition.

Larsen, R. J. and Ketelaar, T. (1991). Personality and susceptibility to positive and negative emotional states. *J Pers Soc Psychol*, 61:132–140, 07.

Nautsch, A., Jasserand, C., Kindt, E., Todisco, M., Trancoso, I., and Evans, N. (2019). The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps Towards a Common Understanding. In *Proc. Interspeech 2019*, pages 3695–3699.

Siegert, I. and Ohnemus, K. (2015). A new dataset of telephone-based human-human call-center interaction with emotional evaluation. In *Proc. of the 1st International Symposion on Companion Technology (ISCT 2015)*, pages 143–148, Ulm, Germany, September.

Siegert, I., V.Silber-Varod, Carmi, N., and Kamocki, P. (2020). Personal data protection and academia: Gdpr issues and multi-modal data-collections "in the wild". *The Online Journal of Applied Knowledge Management: OJAKM*, 8:16 – 31.

Sinha, Y. and Siegert, I. (2022). Performance and quality evaluation of a mcadams speaker anonymization for spontaneous german speech. In *Fortschritte der Akustik - DAGA 2022*, pages 1185–1188, Stuttgart, Germany.

Tomashenko, N., Wang, X., Vincent, E., Patino, J., Srivastava, B. M. L., Noé, P.-G., Nautsch, A., Evans, N., Yamagishi, J., O'Brien, B., et al. (2021). The voiceprivacy 2020 challenge: Results and findings. *arXiv preprint arXiv:2109.00648*.

Watzlawick, P., Beavin, J. H., and Jackson, D. D. (1967). *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies, and Paradoxes*. Norton, Bern, Switzerland.

WP29 (Article 29 Data Protection Working Party). (2013). Opinion 03/2013 on purpose limitation.

WP29 (Article 29 Data Protection Working Party). (2014). Opinion 05/2014 on anonymisation techniques.

WP29 (Article 29 Data Protection Working Party). (2018). Opinion guidelines on transparency under regulation 2016/679, revised.

Zeng, Z., Pantic, M., Roisman, G., and Huang, T. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31:39–58.