The Voice and Speech Processing within Language Technology Applications: Perspective of the Russian Data Protection Law

Ilya Ilin

PhD Student, School of Law, University of Tartu. Address: Näituse 20, 50409 Tartu, Estonia. E-mail: ilya.ilin@ut.ee

Abstract

Language technology (LT) in its broad sense comprises speech technology, computational linguistics, and natural language processing technology. These technologies are expected to have great economic potential and a considerable impact on the everyday life of society. The development of LT fosters applications for artificial intelligence (AI) and broadens the horizon for its advancement. LT deals not only with written forms of linguistic expression but also extends to voice and speech. Voice excluding speech or its contents is a combination of unique physical patterns, such as vocal qualities, volume, speed, and certain other biometric data. Voice can provide medically relevant information, e.g. about a person's mental state, stress level, etc., which is potentially sensitive medical data. Voice with inclusion of speech content can also include personal data (e.g. name, address, ID number, etc.). Consideration of voice and speech as personal data presents a range of legal vulnerabilities and challenges for developing and disseminating LT. This paper explores the extent to which the special regime used for personal data derived from voice and speech affects how it is processed and how it bears on the development and dissemination of LT. This investigation will identify legal vulnerabilities that arise in this connection, and its findings should be useful to both researchers and entrepreneurs in LT. The results of this study provide a basis for further research into LT and related legal issues concerning personal data in Russia.

Keywords

personal data, protection; biometric data; data-intensive product; language data; language technology.

For citation: Ilin I. (2020) The Voice and Speech Processing within Language Technology Application: Perspective of the Russian Data Protection Law. *Legal Issues in the Digital Age,* no 1, pp. 99–123.

Introduction

The rate of growth in language technology (LT) and its popularity indicate both that this field has great economic potential and that it will have a considerable impact on social development. LT in a broad sense comprises computational linguistics, speech technology, and natural language processing technology. The development of these technologies fosters artificial intelligence (AI) applications and broadens the horizon for its advancement. Examples of LT can be found in almost every aspect of our life. These applications vary from grammar checkers and text translators to applications that can control complex machines, synthesize voice, identify people, and communicate with them.

LT deals not only with the written forms of linguistic expression which generally refers to words, but also includes voice and speech as core elements of the communication process. Voice makes the communication process fast and facilitates inputs of data and interaction between computers and people (Holmes W., 2001: 1).

Voice and speech can be used as an element of language data (e.g. vocalized texts, audio records, broadcasts, etc.) for creation of models and datasets or as the input or output for LT products and applications.

The usage of voice and speech within LT requires legal compliance with the regulations that are applicable, and that to a large extent depends on the legal status of voice and speech. The human voice and speech are legally complex phenomena. Voice and speech can be simultaneously covered by copyright, related rights (mainly a performer's rights), rights of the data subject and personality rights. This study focuses on voice and speech from the perspective of Russian law pertaining to data protection by examining the development and dissemination of LT within the legal framework defined by the Russian model of data protection.

In most cases, voice and speech are analyzed together as one complex object. At the same time, one should note that there is a difference between the terms "voice" and "speech". Voice refers to a process that creates acoustic waves. refers to a process that creates phonemes. In other words, it is possible to consider voice as the vocal component of speech (Behrman A., 2017: 4).

Voice without speech and its contents refers to a combination of unique physical patterns such as vocal qualities, volume, speed and certain other

biometric data. Voice can provide medical information, e.g. person's mental state, stress level, etc. and can contain sensitive medical data. Voice in connection with speech content can also include personal data (e.g. name, address, ID number, etc.).

The difference between these two terms should be recognized. When it is essential for analysis, voice and speech may be studied separately from each other.

Consideration of voice and speech as personal data presents a range of legal vulnerabilities and challenges due mainly to the necessity of processing voice and speech for the purpose of developing and disseminating LT. This paper will explore to what extent the special regime for handling personal data affects the development and dissemination of LT, and it will identify and classify the related legal liabilities. The paper should be useful both to researchers and entrepreneurs in LT. The results of this study provide a basis for further research into LT and legal issues concerning personal data in Russia.

The paper is divided into three main sections and a conclusion that summarizes the findings. The first section focuses on the types of personal data with respect to the context of voice and speech processing within LT. The second section analyzes the data protection rules for voice and speech processing. Legal compliance with these rules affects LT development and dissemination. The third section aims to identify the limits of such compliance. The identification of limits is based on legal analysis provided in the previous sections and on the material, temporal and territorial scope of the data protection regulation.

1. Definition of Voice and Speech as They Relate to Data Protection

The right to personal data protection arises from developments in technology. (Hijmans H., 2016: 48) The development of the information and communication (ICT) sector, the increase in cross-border data flows, and the transition to a digital economy have led to problems caused by easy access to personal data (Hungerland F., et al., 2015: 33, 57). In this context, personal data requires a special regime of legal and technical protection. The special legal regime for personal data, on the one hand, ensures protection of the rights belonging to the subject about whom data has been collected. On the other hand, it places a legal restriction on its optimal usage by ICT products.

The first problem in personal data protection is to identify which data is personal. Obviously, such data as names, passport data and addresses are personal. However, determining what is personal may be more involved when it comes to more legally complicated things such as voice and speech. At the same time, consideration of voice and speech as personal data places them under a special legal regime and therefore affects their further processing and use.

There are two general approaches to the analysis of voice and speech with respect to personal data protection (see fig. 1)

According to the first approach, voice and speech are to be regarded as a general category of personal data. The main focus of this approach is on speech content (speech data).

The second approach considers voice without much emphasis on speech data and content. The main focus is on voice and its unique combination of physical patterns that is legally designated as belonging to special categories of personal data (e.g. health data, biometric data).

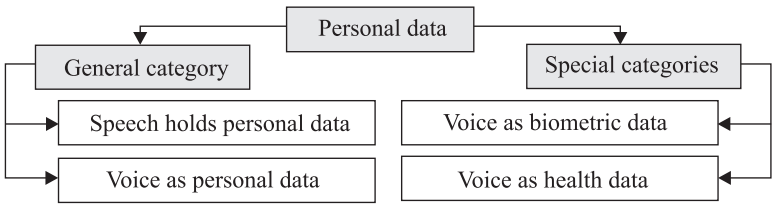

Fig. 1. Voice and speech as personal data

Russian data protection regulations apply to speech processing in the event that the speech data and its content include personal data. The Federal Law "On personal data[1]" defines personal data as any information that refers to an identified or identifiable natural person (data subject[2]). This definition is broad and covers practically any data about individuals. For instance, existing case law has found that the following kinds of data are personal: surname, name and patronymic; year, month, day and place of birth; address; family and social status; property status; education, profession and

[1] Federal Law "On personal data" No. 152-FZ dated 27 July2006, entry into force: 26 January2007. Available at: URL: https://pd.rkn.gov.ru/authority/p146/p164/ (accessed: 18.05.2020). All translations from Russian into English are by the author unless otherwise noted.

[2] Article 11 Federal Law "On personal data" No. 152-FZ.

income[3]; passport data[4], e-mail address[5], and information on crossing national borders[6].

This broad understanding of which data are personal implies that voice and speech should be regarded as personal data whenever they refer to an identified or identifiable data subject.

However, there is still the question of how to apply the data protection regulations when LT developers do not know the identity of the subject whose voice and speech data are being processed within LT. For example, there could be voice samples without any linked descriptions and information. The Federal law "On personal data" does not provide a definite answer to this question. At the same time, the European Court of Human Rights (ECHR), whose case law applies to Russia, does provide protection under those circumstances[7].

The Russian data protection regulations specify three main categories for personal data: general, special and biometric personal data. There is also a fourth category of personal data — publicly available personal data — which was established by Decree of the Government of the Russian Federation No. 1119 "On approval of the requirements for the protection of personal data when processing them in information systems of personal data[8]".

Russian data protection law defines publicly available personal data as data that has been included in publicly accessible sources (directories, ad-

[3] Case law: Presidium of the Russian Supreme Arbitration Court, Resolution in case No. A36-5713 / 2014, dated 29 April 2015, available at: URL: https://kad.arbitr.ru/Card/21af41bd-86ed-4551-b372-10bb6499cf3d (accessed: 18.05.2020)

[4] Case law: Appeal Determination of the Moscow City Court dated 22 May 2014, No. 33-14709, available at: https://mos-gorsud.ru/mgs/services/cases/appeal-civil/details/957f8cd4-63f9-4f26-bfc2-223eec1fb06c?caseNumber=33-14709 (accessed: 18.05.2020)

[5] Case law: Kalininsky District Court (St. Petersburg, Russia), Decision No. 12-253 / 2015 dated 26 May 2015, available at: URL: https://kln--spb.sudrf.ru/modules.php?name=sud_delo&name_op=sf&delo_id=1540005 (accessed: 18.05.2020)

[6] Case law: Moscow City Court, Appeal Determination dated 10.04.2014, No. 33-11688, available at: URL: https://mos-gorsud.ru/mgs/services/cases/appeal-civil/details/9b7aa84e-2dc9-4599-8f70-4edb1a9eb708?caseNumber=33-11688 (accessed: 18.05.2020)

[7] Case law: ECHR. *S. and Marper v. the United Kingdom*, Nos. 30562/04 and 30566/04 [GC], 4 December 2008, § 84. available at: https://rm.coe.int/168067d216 (accessed: 18.05.2020)

[8] Clause 5 of the Decree of the Government of the Russian Federation No. 1119 "On approval of the requirements for the protection of personal data when processing them in information systems of personal data".

dress books[9]) with the explicit consent[10] of the data's subject. The placement of personal data without explicit consent in public sources does not automatically make it publicly available[11]. Publicly available personal data is still considered personal data and should be processed in compliance with data protection regulations[12]. However, there are fewer requirements for processing it. For instance, those data may be processed without consent[13]. Publicly available biometric data, however, is an exception, and it may be processed only with the consent of data subject.

The publicly available category of personal data is excluded from the general tripartite division of the personal data analyzed here for two reasons. First, the Federal law "On personal data" does not classify it as an independent category; and second, it is reasonable to assume that the availability characteristic in general refers to the location and manner of data storage rather than to the characteristics of the data itself.

One special category of personal data is data that indicates political opinions, racial or ethnic origin, philosophical or religious beliefs, and health or sexual orientation[14]. Biometric data are those that refer to the biological and physiological characteristics that can be used to identify a person[15] (e.g. DNA, fingerprints, voiceprints, the image, eyes, body structure[16]).

The tripartite division of personal data into general, special and biometric is the initial prerequisite for data processing. For instance, biometric data can be processed only after the explicit consent of the data subject has been received[17]. Processing of the special category of personal data is generally

[9] Article 8 (1) Federal Law "On personal data" No. 152-FZ.

[10] The data subject has the right to withdraw consent (Article 8 (2) Federal Law "On personal data" No. 152-FZ).

[11] Case law: Decision of the Moscow District Arbitration Court of 09 November.2017 in case No. A40-5250/2017, available at: https://kad.arbitr.ru/Card/eb1907d9-be95-4b0e-85c7-0481aef89b31 (accessed: 18.05.2020)

[12] Article 6 (1) Federal Law "On personal data" No. 152-FZ.

[13] Article 6 (1) Federal Law "On personal data" No. 152-FZ.

[14] Ibid. Article 10.

[15] Ibid. Article 11.

[16] "Explanations of the issues in attributing photo, video, fingerprint data and other information to biometric personal data and the features of their processing" issued by Roskomnadzor on 30 August 2013, available at: URL: http://www.garant.ru/products/ipo/prime/doc/70342932/ (accessed: 18.05.2020)

[17] Article 11 Federal Law "On personal data" No. 152-FZ.

prohibited[18]. In addition, the different categories data require different levels of protection (Krivogin M., 2017: 82–83).

The main criteria used to classify data as personal is the identifiability of a natural person, which to a great extent depends on the context of processing data. Depending on the context, data may be identifiable for one person and not identifiable for others (Oostveen M., 2016: 306).

The context of voice and speech processing within LT is affected by the way it is used and by the technology applied. These factors define the number of activities that may be executed through voice and speech.

Voice and speech can be used in two ways. In the first, voice and speech are considered an input for an existing application (e.g. a voice command made to a voice-operated assistant). The second way is to use voice and speech as language resources (LR) for creating LT applications and to treat them as sources of the language data that they contain,

Creating an LT application largely depends on the existence and number of the LR available (Jents L. and Kelli A., 2014: 164–165). LR are a core element of an LT application and in a broad sense may be described as the range of datasets consisting of texts in oral and written form (language data) which are subsequently used in a machine-learning process (Kelli A. et al., 2018: 79). Creation of LR depends upon two consecutive processes: digitalizsation of language by collecting and transforming the language data into a machine-readable form; and mining texts by analyzing data with a machine-learning algorithm (Jents L. and Kelli A., 2014: 167–170).

These classifications are essential for determining the limits to legal compliance with data protection rules. Those limits are discussed in the third section of this paper.

The context for voice and speech processing within LT is also affected by the technology applied. It could be voice biometrics, speech analysis, speech recognition and speech synthesis. Each type of voice and speech processing focuses on a different kind of information included in voice and speech.

Voice biometrics takes the human voice as a unique personal characteristic that can be used to identify a person along with DNA and fingerprints (Jain A.K., et al., 2004: 4–7). Speech analysis deals with the information which can be obtained by voice, such as level of stress, emotional state,

[18] Ibid. Article 10.

mood and other data concerning a person's mental condition (Chang K., et al., 2011: 1–2). Speech recognition is used to convert speech into text through automatic transcription (Clark A., et al., 2013: 299), and the reverse process is speech synthesis which is used to vocalize a text by converting the text materials into speech (Dutoit T., 1997: 1). Speech synthesis technology does not produce a real human voice that can be recognized and then traced to a particular person. However, that technology is included in this analysis because it is built on neural networks. Neural networks are trained with real examples of human speech (e.g. voice recordings, radio broadcasts), and therefore personal data is still being used in developing of speech synthesis applications (Jents L. and Kelli A., 2014: 172–174). Moreover, personal data could be an output of this kind of technology.

It follows from this description of voice and speech processing by LT that voice and speech can be categorized into the following types of personal data (Table 1).

Table 1

Voice and speech processing and personal data categories

Type of processing	Way used / Information	Personal data category
Voice biometrics	**Input:** special physical characteristics	Biometric data
Speech analysis	**Input:** special physical characteristics	Special data category — Health data
Speech recognition	**Input:** speech content **LR:** Language data for LT creation	General data
Speech transcription	**LR:** Language data for LT creation	General data

Processing voice without a definite connection to speech data and its content should be classified as biometric data. The Russian data protection regulations differentiate biometric data from the other personal data categories. Biometric data reveal the physiological, physical, or behavioral characteristics of a natural person[19]. Voice processing as biometric data in LT has two main purposes: to verify the identity of a person (voice biometrics) or to gain a new piece of information about a person (voice and speech analysis) (Jobanputra N., et al., 2008: 6).

[19] Article 11 Federal Law "On personal data" No. 152-FZ.

Like fingerprints or facial recognition, voice biometrics uses voiceprints as a way to verify and identify a natural person. Biometric systems come in two modes: verification and identification. Verification mode means that a voiceprint is compared with the voiceprint that was originally used to set the identity being claimed. Identification mode means that the system scans the database of voiceprints to find a match, which establishes an identity (Jain A.K., et al., 2004: 1–3). Voiceprints are often used in combination with other categories of personal data. For instance, a bank's voice security system may also ask a client to provide their ID or telephone number. In this scenario, the system checks both the voiceprint and the personal data provided.

The Russian data protection law designates information as biometric data only if the operator uses physiological and biological characteristics for identification purposes[20]. The use of data processing for the purpose of identification is the main characteristic which indicates that a piece of biometric data is personal biometric data[21]. Hence, voice should not be regarded as personal biometric data unless it is used for identification purposes.

Speech analysis processes voice and speech (their characteristics) in order to gain a new piece of information about a person's state. For instance, voice and speech analysis are often used in medical applications. (Chang K., et al., 2011: 1–2) because they can provide data about emotional states, level of stress (Hafen R. and Henry M., 2012: 499–502) or other information concerning health.

At this point it would be natural to ask whether voice should always be considered health-related data or not. Russian data protection regulations do not specify what information is health-related. However, the regulations pertaining to preservation of health do establish the concept of a medical secret and stipulate that all information about requests for medical assistance, information about illnesses, or information obtained through medical treatment and examination should be considered medical secrets[22]. The disclosure and processing of such information are prohibited, although there are a

[20] Ibid.

[21] "Explanations of the issues in attributing photo, video, fingerprint data and other information to biometric personal data and the features of their processing", issued by Roskomnadzor on 30 August 2013, available at: URL: https://pd.rkn.gov.ru/press-service/subject1/news2729/ (accessed: 18.05.2020)

[22] Article 13 Federal law "On the fundamentals of protecting the health of citizens in the Russian Federation" No. 323-FZ, entry into force: 22 November 2011. Available at: URL: http://kremlin.ru/acts/bank/34333 (accessed: 18.05.2020)

few exceptions[23]. The analysis of secret medical data justifies classifying it as a subgroup of the special data category concerning health.

There is no reason to maintain that voice is always health-related data and therefore to provide special legal treatment for it. If there were such a reason, all broadcasting, radio, music and TV shows would have to be classified as processing special data (health data). Voice is properly regarded as health data only when it is intentionally used to obtain information about health.

The analysis of voice and speech as personal data indicates that practical approaches to defining personal data recognize that voice and speech are personal data. It should be noted that there was a case in which recorded voice was not regarded as personal data (Arkhipov V. and Naumov V., 2016: 879). Nevertheless, the common understanding of personal data does not leave much room to argue that voice and speech are not personal data. At the same time, there is still a question about classifying it into an appropriate category of personal data.

Proper definition of the personal data category for voice and speech has important consequences for processing them in LT. Each category has a different level of protection and therefore different regulatory rules for their processing. In the following section, these regulatory rules are analyzed in relation to each respective data protection category. Voice and speech may be classified as in the general personal data category which is covered by general rules of personal data processing and also as in special categories of personal data, such as health and biometric data which have special requirements for their processing.

2. Regulatory Rules for Voice and Speech Processing

Whenever voice and speech are designated as personal data, their processing by LT should be carried out in compliance with data protection rules. Russian data protection regulations define personal data processing so broadly that virtually all manipulations of personal data are included. The Federal law "On personal data" states that processing includes operations with data which are performed by non-automatic or automatic means and are connected with collecting, recording, structuring, storing, usage, transmission and so forth[24].

[23] Ibid. Article 13 (3).

[24] The complete list of the operations that are regarded as data processing is established by Article 3 (3) Federal Law "On personal data" No. 152-FZ.

There are usually several parties engaged in data processing. For instance, the voice identification made by bank security systems involves transfer of the collected voice samples to a voice database that could be in locations remote from the bank. Russian data protection regulations singles out only one entity which can perform data processing (the operator). The Federal law "On personal data" defines the operator as a special entity (a natural or legal person, government authorities) performing data processing and defining its scope, methods and purposes[25]. The operator is the key figure in personal data processing. The technical process of data processing can be arranged by an operator directly or an operator may delegate data processing to a third party[26].

The primary and fundamental principles for personal data processing have been determined by Article 5 of Convention No. 108[27] and are reflected in Article 5 of the Federal law "On personal data". In accordance with Article 5 of Convention No. 108, personal data is to be processed and collected lawfully[28] and fairly[29]; the data must be relevant[30]; processing must be limited to the purposes for which it was stored[31], accurate[32] and kept in a form which allows identification of the data subject no longer than required for

[25] Ibid. Article 3 (2).

[26] Ibid. Article 6 (3).

[27] Article 5 of the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, reference ETS No.108, treaty open for signature by the member States of the Council of Europe and for accession by the European Union at Strasbourg 28 January 1981. Entry into force: 1 October 1985, available at: https://www.coe.int/en/web/conventions/full-list/-/conventions/treaty/108 (accessed: 18.05.2020)

[28] Article 5 (a), Article 5 (b) Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, reference ETS No.108. Case law: ECHR, *Taylor-Sabori v. the United Kingdom* No. 47114/99 22 October 2002, available at: http://hudoc.echr.coe.int/eng?i=001-60696 (accessed: 18.05.2020); ECHR, *Peck v. the United Kingdom* No.44647/98 28 January 2003, available at: http://hudoc.echr.coe.int/eng?i=001-60898 (accessed: 18.05.2020); ECHR, Khelili *v*. Sweden, No, 16188/07, available at: http://hudoc.echr.coe.int/eng?i=001-107033 (accessed: 18.05.2020)

[29] Article 5 (a) Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, reference ETS No.108. Case law: ECHR, *Haralambie v. Romania*, No 21737/03, 29 October 2009, available at: http://hudoc.echr.coe.int/eng?i=001-95397 (accessed: 18.05.2020); ECHR, *K.H. and others v. Slovakia*, No.32881/04 28 April 2009, available at: http://hudoc.echr.coe.int/eng?i=001-92418 (accessed: 18.05.2020)

[30] Article 5 (c) Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, reference ETS No. 108.

[31] Ibid. Article 5 (b).

[32] Ibid. Article 5 (d).

the purpose of storing the data[33]. These are the main principles of personal data processing for guaranteeing a minimum level of the protection for it. Additional rules for personal data processing are based on these fundamentals.

The data protection rules fall into three groups: rules concerning security of processing, the lawfulness of processing, and transparency of processing. Voice and speech may come under the special and biometric personal data category and therefore be classified as sensitive data; or they may be in the general personal data category and therefore be treated as non-sensitive data. The legal framework for processing of these two categories should be examined with this in mind.

The first group of rules stipulates security measures that should be applied in data processing. Under the Russian data protection regulations, these measures should be implemented by the operator engaged in personal data processing. There are two groups of security measures: technical and organizational[34]. The Federal law "On personal data" provides only general provisions for the security measures. In practice, the operator in personal data processing is to arrange for an audit of the information systems that are used for personal data processing and identify which of the four categories is applicable to the systems[35]. Proper identification of an information system's category is crucial for assigning the level of threat and determining security measures[36].

The second group of data processing rules is derived from the principle of lawfulness. This principle presumes that personal data processing must be executed in strict compliance with the law and be legally justified.

Russian data protection regulations allow the following grounds for non-sensitive personal data processing: consent of the data subject; contractual performance; compliance with a legal obligation; protection of vital interests; performance of a task carried out in the public interest; and processing

[33] Ibid. Article 5 (e).

[34] Article 19 Federal Law "On personal data" No. 152-FZ.

[35] According to the Order of the FSTEC of Russia, the Federal Security Service of Russia, and the Ministry of Information Technologies and Communications of Russia No. 55/86/20, 13 February 2008, four classes of information systems exist.

[36] Decree of the Government of the Russian Federation 1 November 2012 No.1119 "On the approval of the requirements for the protection of personal data when processing them in information systems of personal data", available at: URL: https://rg.ru/2012/11/07/pers-dannye-dok.html (accessed: 18.05.2020)

for legitimate interests[37]. Moreover, non-sensitive personal data can be processed for statistical reasons[38] or processing may be done in order to comply with an obligation to disclose information[39].

The rules for processing sensitive personal data vary depending on its data protection category. Hence, the rules are different for voice and speech processing when they are processed as either health or personal biometric data.

Processing of health data is in general prohibited[40]. However, there is no blanket restriction on biometric data processing, which may be performed with consent from the data subject[41].

An analysis of the existing justifications for lawful personal data processing yields the conclusion that the most pertinent legal grounds for voice and speech processing by LT are consent and the legitimate interest. However, if an LT has been developed by research units, the legal justification of performing a task in the public interest by conducting research is applicable as well.

The last group of rules for personal data processing concern transparency in data processing. The transparency of data processing is defined as the data subject's right to ascertain the existence of automated personal data processing, its main purposes, and the identity and habitual residence or place of business of the controller of data processing operations[42].

These principles and rules for personal data processing should be applicable to voice and speech processing by LT under the appropriate personal data category. Compliance with these rules establishes the scope of a data subject's legal rights concerning personal data protection.

However, there is still the question of the limits of this compliance. In other words, does the application of data protection rules extend to voice and speech processing? Furthermore, it should be acknowledged that a data subject's rights are not absolute and that they should be weighed together with other fundamental rights such as freedom of thought, expression and infor-

[37] Article 6 Federal Law "On personal data" No. 152-FZ.

[38] Ibid. Article 6 (1–9).

[39] Ibid. Article 6 (1–11).

[40] Article 10 (1) Federal Law "On personal data". A list of the exceptions to the general rule is provided in: ibid. Article 10 (2).

[41] Ibid. Article 11.

[42] Article 8 (a) Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, reference ETS No.108.

mation, and the right to linguistic and religious diversity (Docksey C., 2016: 197–199) In the next section, the limits on this compliance are investigated.

3. The limits of Compliance with Data Protection Rules

The processing of voice and speech by LT should be carried out in accordance with data protection rules. However, to what extent does data protection regulation apply to voice and speech processing? For instance, suppose that a language model for natural language processing has been created by using personal data. Does that mean that further use of the products based on that model should be subject to data protection regulations?

The limits of data protection regulations can be established by reference to the material, time and territorial scope of the data protection regulations concerning voice and speech processing by LT.

The material scope of data protection regulations pertaining to voice and speech processing can be identified with the various levels involved in LT product development. Those levels include collecting language data for datasets, compiling datasets, annotation of datasets, models, and creation of a product (Kelli A., et al., publication pending).

Collecting language data is one part of the process of creating LR. Voice and speech are used as raw material in the collection stage, and processing them involves only collection of data along with minor technical manipulations of it. Up to this point personal data cannot be anonymized to such an extent that a data subject cannot be identified.

The next level requires that the language data collected be systemized and organized according to specific conditions. However, the language data remains the same as before, and processing modifies only the systematization of the data. As regards data protection, there is not much difference in the legal status of voice and speech processing at the first and second processing levels. However, there is a technical difference in that it becomes difficult for a data subject to identify that their data has been included in the dataset because of the integrated character of the database (dataset).

The creation of the annotated datasets is the third process in collecting and organizing data. The legal status of voice and speech within datasets are the same as at the previous levels. It should be noted that data annotation occurs according to three scenarios for data analysis — automatic, semi-automatic, or physical — and this bears on issues concerning copyright and

identification of an author for such annotations. But the topic of the intellectual property protection for language data is outside the scope of this paper.

Data collection, systematization and annotation all regard voice and speech only as language data without any consideration of biometrics. The data used is not anonymized, and therefore the main concern is that speech will contain personal data. In this sense, the processing of voice and speech at these levels requires compliance with data protection regulations. This processing should be conducted with the legal justifications appropriate for the general data processing category.

The output of collecting, systematizing and annotating language data are various language datasets such as Open Subtitles[43], the Common Crawl dataset[44], the Universal Dependencies treebanks[45], etc. Some of these datasets are employed subsequently for creating language models that describe the rules for a given language and how that language works. In a broad sense, these examples of models may include pre-training language models (Devlin J. et al., 2018: 1–2), various word lists, n-gram lists, dictionaries, and pre-training word embeddings (Grave E. et al., 2018: 1–2, 5). LT relies heavily on models of this kind as the basis for most LT applications.

Considered as personal data, voice and speech in language models are used as general data, and there is no focus on their unique patterns. Therefore, they cannot usually be placed in the special or biometric data protection category, and this is because a language model incorporates only the general category of personal data (e.g. a voice sample concerning the data subject's name or e-mail). The legal liability in the use of such a model can be minimized by anonymizing the personal data Anonymized personal data as understood in Russian data protection regulations are personal data that do not require identification[46]. The processing of the anonymized personal data is subject to fewer requirements (Mavrinskaya T.V. et al., 2017). However, if the data were not non-personal from day one of its collection and were not anonymized throughout their processing, then the anonymization process is nevertheless classified as personal data processing[47].

[43] Available at: https://www.opensubtitles.org/ru (accessed: 18.05.2020)

[44] Available at: http://commoncrawl.org/ (accessed: 18.05.2020)

[45] Available at: https://universaldependencies.org/ (accessed: 18.05.2020)

[46] Article 3 (9) Federal Law "On personal data".

[47] Ibid. Article 3 (3) states that personal data processing is any action (operation) or a combination of actions (operations) performed both automatically and manually with personal data, includ-

The legal handling of language data does not always correspond to the legal handling of the language model that was built on that data. (Kelli A., et al., publication pending) A language model consists of language rules, and it is a very challenging technical task to extract personal data from the model. Even if a model has been built upon language data that contained personal data, the identifiability of data subjects in most cases is lost once the data has been processed.

However, a question remains about the appropriate use of datasets that contain personal data for creating language models. Because these datasets contain personal data, processing then should be undertaken on proper legal grounds[48]. Choosing the proper legal basis for data processing would depend on the way in which the model will be used. The kind of problem that may arise is illustrated by the following scenario. Suppose that a model has been designed for use in research; the personal data collected has therefore been processed as qualifying for the exemption from restrictions on personal data processing when that data is used for research or as having the appropriate consent. But then suppose that it has been disseminated or made available publicly. In that case, the data could be anonymized, or additional consent that covers commercial use and public access should be obtained. Resorting to these solutions may require substantial technical and procedural adjustments.

Creation of a language model can be considered as the stage after which language data is excluded from the end product (i.e. an LT application). For instance, personal data regulations will not cover a synthesized voice (an output of an LT application), although it has been created by using a language model that included personal data. The legal regulatory status of the used language data does not extend to the end product. Therefore, for the purpose of data protection, the language data regulations no longer apply to LT after a model has been created. However, the legal status of the inputs (e.g. voice commands) should still be ascertained by considering personal data protection.

The time limits for data protection are determined by the duration of data protection rights, and it is therefore crucial to establish when the data subject's rights expire. For instance, there was a case in which the Russian voice

ing collection, recording, arrangement, accumulation, storage, specification (updating, changing), extraction, use, distribution (including transfer), anonymizing, blocking and destruction of personal data.

[48] Ibid. Article 6.

company STC Group synthesized the voice of a dead Russian actor and then vocalized a novel with the synthesized voice[49]. Russian data protection regulations protect the personal data deceased persons[50], and the data processing must be carried out in compliance with data protection rules[51]. At the same time, Russian data protection regulations do not establish the duration of that protection. To fill in this gap, it would be reasonable to make the duration equal to that for protection of a person's private life (Vazhorova M.A., 2012: 57–59). That protection persists for at least 75 years after a person's death[52].

Another concern regarding the limits of data protection regulation is its territorial extent and the external effect of such rules. The problem is that LT products are not usually intended for only one country and are often distributed in different jurisdictions. For instance, the speech-to-text system developed by Google[53] supports more than 120 languages and can be integrated with other ICT products developed in various countries with different models of data protection. The limits for compliance with data protection would then also be determined within the national jurisdictions of the countries to which the LT products are distributed. Do LT developers therefore need to comply with the data protection rules applied where their products are distributed? The situation becomes even more complicated when the LT developers use cloud computing which depends upon trans-border data flows. For instance, the speech-to-text system developed by Yandex[54] is distributed as a cloud service. The Yandex cloud is certified as an information system that fully meets Russian data protection requirements[55]. However, in the

[49] An example of synthesised voice is available at: https://www.youtube.com/watch?v= hvaB1exK9rY (accessed: 18.05.2020)

[50] Case law: Decree of the Federal Arbitration Court of the Eastern Siberian District dated 1 July 2008 No. A33-14182/2007, available at: URL: https://kad.arbitr.ru/Card/c7241b92-6ff6-42ee-b233-b398a3080b4b (accessed: 18.05.2020)

[51] If a personal data subject has died, consent for processing their personal data is to be provided by the heirs of the personal data subject, unless the personal data subject gave such consent while still alive. Article 9 (7) Federal Law "On personal data".

[52] Article 152.2 (5) The Civil Code of the Russian Federation (Part I of IV) No. 51-FZ dated 30 November 1994, entry into force: 1 January 1995. Available at: URL: http://www.wipo.int/edocs/lexdocs/laws/en/ru/ru083en.pdf (accessed: 18.05.2020)

[53] Cloud Speech API, available at: https://cloud.google.com/speech-to-text (accessed: 18.05.2020)

[54] Yandex SpeechKit, available at: URL: https://cloud.yandex.ru/services/speechkit (accessed: 18.05.2020)

[55] Available at: URL: https://storage.yandexcloud.net/yc-compliance/conformance_ru_pdp.pdf (accessed: 18.05.2020)

event that this system is integrated into a European ICT product, the problem of complying with both sets of regulations arises as does the issue of the applicability of data protection laws from different jurisdictions.

Russian national data protection regulation as a general rule does not have an extraterritorial effect. Therefore, it does not apply to non-residents that are processing the personal data of Russian citizens abroad. This rule has two exceptions. The first one concerns the data localization requirement, and the second is a consequence of the anti-terrorism measures addressed in the "Yarovaya package[56]".

The localization rule for personal data of Russian citizens was stipulated for data protection regulation by Federal Law 242-FZ dated 27 April 2017[57]. This amendment created a new requirement that data processing operators store, collect and use personal data of Russian citizens only in databases located within Russian territory[58].

The economic impact of the Russian data localization rule has been studied by the European Centre for International Political Economy (ECIPE). According to the Centre's report, the rule mostly harms the economy and reduces the productivity of Russian companies because they must build their data centers in Russia, and they are not allowed to use similar services abroad (even if it were economically feasible). The ECIPE estimate that the resulting economic losses amount to around 0.27% of gross domestic product[59].

[56] Unofficial named after Irina Yarovaya, one of its authors, the package consists of two Federal Laws: (i) Federal law "On amendments to the Federal Law 'On counteracting terrorism' and certain legislative acts of the Russian Federation regarding the establishment of additional measures to counter terrorism and ensure public safety") No. 374-FZ dated 6 July 2016, entry into force: 20 July 2016. Available at: URL: http://kremlin.ru/acts/bank/41108 (accessed: 18.05.2020); (ii) Federal law "On Amendments to the Criminal Code of the Russian Federation and the Code of Criminal Procedure of the Russian Federation with regard to the establishment of additional measures to counter terrorism and ensure public safety" No. 375-FZ dated 6 July 2016, entry into force: 20 July 2016. Available at: URL: http://kremlin.ru/acts/bank/41113 (accessed: 18.05.2020)

[57] Federal Law "On amendments to certain legislative acts of the Russian Federation regarding the clarification of the procedure for processing personal data in information and telecommunication networks" No. 242-FZ dated 21 July 2014, entry into force1 September 2015. Available at: URL: http://www.consultant.ru/document/cons_doc_LAW_165838/ (accessed: 18.05.2020)

[58] Article 18 (5) Federal Law "On personal data".

[59] Available at: URL: http://ecipe.org/publications/data-localisation-russia-self-imposed-sanction/?chapter=5 (accessed: 18.05.2020)

There are four conditions to be met in order for the Russian data protection rule to apply. The first condition is that the information must contain personal data. Second, this personal data must have been collected (the operator must have obtained these data from third parties). Third, the data must have been processed in a way arranged by an operator. The last condition is that this data must be connected with Russian citizens (Savelyev A., 2016: 144–145).

That fourth condition leads to the problem of determining citizenship within ICT technologies. For example, how can the citizenship be identified for a person who gives a command through voice assistance, or how can the citizenship of a person whose voiceprint is processed be identified? Roskomnadzor (the Russian data protection authority) has issued an official opinion[60] that partly solves this problem. According to this opinion, the term "citizenship" is to be replaced with the territory in which processing takes place. If there are uncertainties about the data subject's citizenship, all information processed and collected within Russian territory must be localized at databases located in Russia[61]. However, it is still unclear how to identify and process personal data of Russian citizens that are collected outside Russian jurisdiction.

The localization rule is crucial for the companies that use cloud services localized in other jurisdictions as well for the companies that provide services in the Russian market, even if they do so without having any branches or representatives within Russian territory. For instance, the social network LinkedIn developed by LinkedIn Corporation has no representative offices, departments or other legal entities in Russia. However, because the company breached the localization rule by processing the personal data of Russian citizens outside of Russian jurisdiction, LinkedIn was banned in Russia[62].

In addition to the localization rule, there is one more exception to the territorial reach of Russian data protection. This exception is also connected with the Yarovaya package, although it is not directly concerned with data protection. It has a different material scope than the Federal law "On personal data" and mostly concerns the public sector (national and public secu-

[60] Letter by Roskomnadzor No. 08АП-3572 dated 19 January 2015.

[61] Letter by Roskomnadzor, p. 5.

[62] Case law: *LinkedIn Corporation v. Roskomnadzor* 02-3491/2016, decision of the Tagansky District Court (Moscow, Russia) dated 4 August.2016; appeals determination of the Moscow City Court dated 10 November 2016 case No. 33-38783 / 16. Available at: URL: https://www.mos-gorsud.ru (accessed: 18.05.2020)

rity). The Yarovaya package introduced special anti-terrorism measures that also created new obligations for data storage and data processing.

The measures it introduced require the organizers of information dissemination and telecommunication service providers to store internet traffic (voice and text messages, photos, videos, sounds, file metadata) for periods from six months to three years. The law also requires that, upon issuance of a special order, encryption keys for decrypting internet traffic be provdided in the event that the required data is stored or processed in encrypted form[63].

This package was adopted in 2016; however, some of the issues it raised are still surrounded by legal uncertainties. For instance, it refers to the concept of "organizer of information dissemination", and the law does provide a legal definition of that entity[64]. However, legal analysis of it shows that it is too broad and may cover every internet service and any webpage that somehow interacts with a user (e.g., placing cookies). The definition of the "organizer of information dissemination" is not limited to any national boundaries and therefore may refer to such internet giants as Google, Facebook as well as to other messenger and communication services, such as WhatsApp, Viber, Skype and Telegram and even to blog owners and blog hosting platforms such as Tumblr, Wix and Medium, to administrators for domain names, etc. This legal uncertainty exposes foreign companies to the legal vulnerability of being considered by government authorities as organizers of information dissemination, and that would make it necessary for these companies to comply with the rules described above.

Complying with those rules, however, can be difficult for companies because they would be forced to violate their own data protection rules (e.g. rules established by the General Data Protection Regulation [GDPR[65]]) or

[63] Article 10.1 Federal Law "on information, information technologies and protection of information" No. 149-FZ dated 27 July.2006, entry into force: 26 January 2007. Unofficial English translation available at: http://www.wipo.int/wipolex/ru/details.jsp?id=15688 (accessed: 18.05.2020); Article 46 (1), Article 64 Federal law "On communications" No. 126-FZ dated 7 July 2003, entry into force: 1 January 2004. Available at: http://www.wipo.int/wipolex/en/details.jsp?id=17111 (accessed: 18.05.2020)

[64] Article 10.1 Federal Law "On information, information technologies and protection of information" No. 149-FZ.

[65] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), dated 27 April 2016, Entry into force: 25 May 2018, available at: https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed: 18.05.2020)

their contractual obligations (e.g. confidentiality clauses). One of the most consequential examples of the impact of the Yarovaya law package on data protection regulation is the Telegram lawsuit[66] that resulted in Telegram being blocked in Russia[67].

The final problem with the territorial scope of the data protection regulations concerns trans-border data flows and cloud computing. For example, most voice assistants provide their services through cloud computing technology.

For that purpose, it is crucial to identify the country, where personal data was collected and compare its national data protection rules with the Russian ones. The possibility of working with trans-border data can be judged only after making those comparisons.

For instance, legally transferring personal data between Russia and European countries currently is complicated. Even though the Russian and European legislation accept similar international legal grounds for processing personal data and they follow the same data protection principles, their laws have not been harmonized, and their different models for data protection are being applied. Most of the concerns are about how the Russian localization requirement and the requirements of the Yarovaya package relate to the GDPR.

One should note that Roskomnadzor attempted to solve the problem with a localization rule regarding trans-border data flows and stated that the personal data of Russian citizens should be initially collected and stored in databases that are located in Russia. However, it can subsequently be copied and transferred to databases located in other countries[68]. However, the problem of harmonizing the rules in the Yarovaya package with data protection regulation has not been solved.

Conclusion

Voice and speech processing by LT in most cases is regarded as processing of personal data. There are not a great many concerns about the clas-

[66] Case law: Case 02-1779/2018. Tagansky District Court (Moscow, Russia), available at: https://mos-gorsud.ru/rs/taganskij/services/cases/civil/details/2cc72aea-39e7-4f8e-adc9-37d170966efa?caseNumber=02-1779/2018 (accessed: 18.05.2020)

[67] Available at: URL: https://www.nytimes.com/2018/04/13/world/europe/russia-telegram-encryption.html (accessed: 18.05.2020)

[68] Letter by Roskomnadzor.

sification of voice and speech as personal data. However, disputes may arise about which category of personal data covers voice and speech.

Depending on the context of their processing, voice and speech may belong to the general data protection category or to the special (health) or biometric personal data category.

Voice and speech are classified as in the general category when they identify a person (the data subject). This could occur when speech contains some personal data or when a voice sample is linked with information that may disclose a particular person's identity.

Voice and speech are classified as health data when the processing is intended to extract information about emotional state, level of stress or other information concerning health.

Voice and speech are classified as biometric data when they are used in biometric systems for personal verification or identification by analyzing unique vocal patterns.

Each category of personal data comes with different rules for voice and speech processing. Hence, the main risk and legal liability for voice and speech processing is brought about by incorrect determination of personal data categories.

There are two approaches to determining the category of personal data for voice and speech. The first approach presupposes that voice and speech are used as language data (a language resource) for creating a language model. In most cases, these models may include data from the general personal data category and only rarely use sensitive personal data. Because a language model does not use voice and speech for verification and identification, it can be assumed that the biometric personal data categories do not apply to language models nor to the data which was used for their creation.

The second approach presupposes that voice and speech are used as an input to LT end products. What kind of language data were used for creating a product is of no importance for this approach, and the emphasis is on which data category is used to make an LT application work. Depending on the technology used in an application and its functions, these data could be classified as either in the general or special categories of personal data.

Classification of voice and speech as personal data requires LT developers to comply with data protection rules, and any processing of voice and speech should be conducted in accordance with data protection regulations.

The limits to that compliance are defined by the material, time and territorial scope of the data protection regulations pertaining to voice and speech processing. The material scope of data protection regulation varies with the stages in the development of an LT product. The need for legal compliance with data regulations applicable to language data ends once the language model has been created. The processing of voice and speech within end products should be carried out in accordance with the data protection rules applicable to the particular category of personal data.

The time limits for compliance with the data protection regulations are governed by the duration of data protection rights. Russian data protection regulations protect the personal data of deceased persons; however, the duration of such protection is not clear. By analogy with the protection of a person's private life, the author concludes that the period of protection should be at least 75 years after a person's death.

The territorial limits of compliance depend on the applicable data protection regulation. There is no uncertainty about the need for voice and speech processing in applications developed and disseminated within Russian territory to comply with the national Russian data protection regulations. However, the situation becomes more intricate when these activities are performed by a foreign company. The existing legal uncertainty in Russian data protection regulation makes the compatibility of Russian data protection rules with different legal systems (e.g. the one) problematic. The existing regulations on data protection mean that foreign LT developers must comply with both their own national data protection rules and with the Russian ones. Hence, companies may find that they must choose which regulation they will breach. The comparison of Russian data protection regulation as it applies to LT with that of other jurisdictions is a matter for further investigation.

References

Arkhipov V. and Naumov V. (2016) The legal definition of personal data in the regulatory environment of the Russian Federation: Between formal certainty and technological development. *Computer Law & Security Review*, no 6, pp. 868–887.

Behrman A. (2017) *Speech and voice science*. San Diego: Plural publishing, p. 482.

Chang K. et al. (2011) AMMON: A speech analysis library for analyzing affect, stress, and mental health on mobile phones. *Proceedings of Phone Sense*. Available at: http://people.eecs.berkeley.edu/~jfc/papers/11/AMMON_phonesense.pdf (accessed: 18.05.2020)

Clark A., Fox C., and Lappin S. (2013) *The handbook of computational linguistics and natural language processing*. Oxford: Wiley, p. 650.

Devlin J. et al. (2018) BERT: Pre-training of deep bidirectional transformers for language understanding. Available at: https://arxiv.org/pdf/1810.04805.pdf?source=post_elevate_sequence_page (accessed: 18.05.2020)

Docksey C. (2016) Four fundamental rights: Finding the balance, *International Data Privacy Law*, no 3, pp. 195–209.

Dutoit T. (1997) *An introduction to text-to-speech synthesis*. Dordrecht: Springer Science & Business Media, p. 275.

Furey E. & Blue J. (2018) She Knows Too Much — Voice Command Devices and Privacy. *29th Irish Signals and Systems Conference (ISSC)*, The Institute of Electrical and Electronics Engineers (IEEE), pp. 1–6.

Grave E. et al (2018) Learning word vectors for 157 languages. Available at: https://arxiv.org/pdf/1802.06893 (accessed: 18.05.2020)

Hafen R. & Henry M. (2012) Speech information retrieval: A review. *Multimedia systems*, no 6, pp. 499–518.

Hijmans H. (2016) *The European Union as guardian of internet privacy*. Cham: Springer International, p. 564.

Holmes W. (2001) *Speech synthesis and recognition*. London: Taylor & Francis, p. 298.

Hungerland J. et al (2015) The digital economy.Strategy 2030 — Wealth and Life in the Next Generation. Berenberg Bank und Hamburgisches Welt WirtschaftsInstitut. Available at: http://hdl.handle.net/10419/121322 (accessed: 18.05.2020)

Jain A. et al (2004) An Introduction to Biometric Recognition. *The Institute of Electrical and Electronics Engineers (IEEE) Transactions on Circuits and Systems for Video Technology*, no1, pp. 4–20.

Jents L. & Kelli A. (2014) Legal aspects of processing personal data in development and use of digital language resources: the Estonian perspective. *Jurisprudencija*, no 1, pp. 164–184.

Jobanputra N. et al (2008) Emerging security technologies for mobile user accesses. *The electronic Journal on E-Commerce Tools and Applications*. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.374.5082&rep=rep1&type=pdf (accessed: 18.05.2020)

Kelli A. et al (2018) Processing personal data without the consent of the data subject for the development and use of language resources. Selected papers from CLARIN Annual Conference, 2018. Linköping University Electronic Press. Available at: https://www.ep.liu.se/ecp/159/008/ecp18159008.pdf (accessed: 18.05.2020)

Kelli A. et al (2012) Copyright and Constitutional Aspects of Digital Language Resources. *Juridica International*, vol. 19, pp. 40–48.

Krivogin M. (2017) Peculiarities of Legal Regulating Biometric Personal Data. *Law. Journal of the Higher School of Economics*, no 2, pp. 80–89 (in Russian)

Mavrinskaya T.V. et al (2017) Anonymizing personal data and Big Data technology. *Interactivnaya nauka*, no 16, pp. 1–8 (in Russian)

Oostveen M. (2016) Identifiability and the applicability of data protection to big data. *International Data Privacy Law*, no 4, pp. 299–309.

Savelyev A. (2016) Russia's new personal data localization regulations: A step forward or a self-imposed sanction? *Computer Law & Security Review*, no 1, pp. 128–145.

Soldatova V.I. (2020) Protection of personal data in applying digital technology. *Lex Russica*, no 2, pp. 33–43 (in Russian)

Vazhorova M.A. (2012) The relationship between the concepts of "information about private life" and "personal data". *Bulletin of the Saratov State Law Academy*, no 4, pp. 55–59 (in Russian)