Voicing Concerns: The balance between data protection principles and research developments in forensic speech science

Abstract

The status of forensic speech recordings among existing data protection guidance is not clear. The inherent nature of voice and the way in which forensic speech casework is currently allocated mean that there are additional barriers to incorporating real casework data into research activities. The key objective of this work is to explore data protection solutions that could enable the forensic speech science community to responsibly use real casework data for research and development purposes. While reviewing relevant guidance and rulings, issues such as proportionality, opportunism and data minimisation are addressed, as well as where voice sits in relation to the definition of "biometric data". This paper ultimately places forensic speech recordings in the data protection context to illuminate the specific issues that arise for this data type.

Keywords: data protection, forensic speech recordings, data retention, proportionality

1. Introduction

Forensic speech science is the forensic subdiscipline concerned with analysing speech recordings when they arise as evidence. A forensic speech scientist or forensic phonetician may be enlisted to analyse recordings in order to address questions surrounding the identity of speakers. Often, the task will involve analysing the speech of an unknown speaker in an incriminating telephone call and comparing it to the speech of a suspect who has been recorded during a police interview. Like other forensic subdisciplines, forensic speech science is working on advancing analysis methods (including incorporating automatic speaker recognition systems to complement the human expert analysis of voices) and working on implementing recognised scientific quality standards. In the UK, the Forensic Science Regulator oversees the provisions of forensic science services which involves, among a number of things, a focus on compliance with the relevant quality standards. As part of the role, the regulator encourages a shared understanding of quality and standards among all the stakeholders within the Criminal Justice System. This includes the recognition that forensic science needs to be supported by ongoing research in order to maintain and increase quality and capability. Inevitably, there is great scope and need for further research and innovation in forensic speech science, but there are shortcomings to the resources that we currently have at our disposal.

The assertion put forward in this paper is that real casework data is a key component to making meaningful research developments in forensic speech science. Of course, this is not to minimise the efforts by members of the community to create “casework-like” data. Producing replications of casework-like data for research purposes is a well-established approach within forensic speech science, and these datasets have contributed towards valuable work. Such endeavours began with the *Dynamic Variability in Speech* (DyViS) corpus [1], where the authors recruited over one hundred young male speakers of Standard Southern British English to take part in a mock criminal scenario. Here, they were recorded under forensically relevant conditions (i.e. during a telephone call with an accomplice and during a mock police interview). Further corpora have since been created in a similar way to represent the accents in different parts of the UK: *The Use and Utility of Localised Speech* (TUULS) corpus [2] which reflects accent varieties in the North East of England and the *West Yorkshire Regional English Database* (WYRED) [3] which reflects accent varieties in West Yorkshire. All of these forensic speech dataset projects coincide with the UK Forensic Science Regulator’s suggestion that digital forensic disciplines (such as forensic speech science) can “generate effective and comprehensive test data” for research purposes (in the *Forensic Science Regulator’s Protocol* for validation using casework material ([4], p 36)). This supposedly contrasts with other forensic disciplines, such as physical or biological evidence (e.g. blood splatter analysis where it may be assumed that it is more difficult to create such test data). In view of these assumptions, the Regulator suggests that in order to carry out forensic speech science research ‘the need for casework material is … less likely’ ([4], p 36). We propose that it would be a mistake to assume that an area like forensic speech science can easily create forensically-realistic test data.

While existing research datasets, such as those listed above, manage to capture some of the conditions of casework, we would never be able to capture the genuine pressures of forensically realistic environments and the effects that these have on the speech produced. For example, it is not possible to reproduce the emotional impact, the high-stakes situations, the lengths of time that pass between the recordings being compared, and other associated factors that influence speech production. One rare example of work that aimed to look into these sorts of extreme influences on speech was [5] who investigated the effects of distress on speech production and perception. As part of her work, [5] analysed and compared speech recordings of genuine victims requesting assistance after a violent attack against speech recordings of actors pretending to be victims. Within this work, [5] points out how difficult it is to obtain distressed speech recordings, partly as these are calls of a sensitive nature, but also because of the ethical barriers involved in eliciting genuinely distressed speech from volunteers ([5], p 6).

It is also unfeasible to cater for the full range of possible combinations of casework environments through manufactured speech datasets (e.g. indoors, outdoors, varying levels of background noise, different distances to the microphone). Currently, we rely on the training and experience of the forensic practitioners to apply their expertise in order to recontextualise the findings of research carried out on experimental data to casework material. However, if we are committed to identifying the best approaches for analysing and interpreting real casework data, then we need to bring real casework data into the research environment.

In further support to the point that such data replications are compromises in forensic speech science, these corpora have also been known to fall short in court. The third author recalls an instance of when she has referred to findings that had been generated on the DyViS corpus in her forensic speech analysis evidence. The discrepancy between the data in the DyViS corpus and the type of data and conditions in the specific case was highlighted by the barrister. It is accepted that these casework-like corpora enable widespread research, but the findings or outcomes of the research will only go so far if they are not also tested on real case data.

Using real casework data for forensic speech science research is not necessarily impossible (indeed, other forensic disciplines do it), but a key barrier to using real forensic voice recordings is the lack of clarity around how we should treat forensic voice data with regards to data protection. Within forensic science more broadly, there is a general recognition that real case data can be valuable in advancing analysis methods. There have been (and still are) active efforts to develop regulation and guidance on the storage and use of such data (and Sections 2 and 3 of this paper discuss these efforts). Much of the existing guidance, however, aims to account for a broad array of forensic data types, and does not accommodate the special case of forensic voice data, nor the current position of the forensic speech science field[1]. This paper therefore focusses the discussion on forensic voice data. To do so, we will address the following two objectives:

1. to navigate through relevant discussion surrounding data protection and to raise issues that are specifically attached to the protection of forensic voice data;

2. to highlight key ways in which forensic speech analysis providers can align with existing data protection principles and recommendations.

In carrying out this exploration, we seek to contribute towards the longer-term objective of enabling academics and practitioners in forensic speech science to responsibly use real casework material for research purposes. Section 2 of this paper will first outline and discuss some general principles of data protection and where forensic data sits among these, referring to the allowances made for law enforcement purposes. Following this, Section 3 evaluates the different perspectives of relevant bodies and the public, and what these might mean for forensic voice data. The priorities and perspectives of these different bodies do not neatly align, but themes emerge in relation to proportionality of data retention and use, as well as the implementation of a discriminatory approach to data retention. In light of these discussions, Section 4 moves on to suggest practical measures that forensic speech practitioners can put in place in order to construct the environment and processes required to responsibly use casework material for research. Section 5 foregrounds some of the key points and contains final reflections. While this paper specifically considers voice data, we very much suspect that the points raised will apply to other forensic disciplines.

2. Data Protection

When creating solutions for responsibly storing and using data, it is important to remind ourselves of why we need to put safeguards in place at all. Personal data links to an individual's fundamental right to privacy [6]. Beyond fundamental rights, there are also practical risks attached to the existence of personal data. For example, having access to

[1] That said, it is also not the case that there is a great deal of clarity for many other biometric and forensic data types. In recognition of this, the Ada Lovelace Institute has recently commissioned the Ryder Review which will independently evaluate the current regulatory framework (or lack of one) that covers a range of biometric data types (URL: https://www.adalovelaceinstitute.org/project/independent-review-governance-of-biometric-data-uk/ accessed:06/04/2021).

another person’s data can open up the possibility of carrying out fraud. Data theft is a very real risk whether the data is digitally stored or not, but the digital age has led to an increase in opportunities for data theft and subsequent fraudulent or unintended use of it. We are now able to store more data, and analyse more data, but it is now also possible to “steal” and “leak” greater volumes of data. In 2020, it was reported that a cyber-attack on EasyJet resulted in the contact and credit card details of thousands of customers being stolen [7]. Also in 2020, it was reported that a simple security misconfiguration meant that the personal details of millions of Microsoft customers were left on a server where a password was not needed to access them [8]. Anyone with an internet connection could have obtained these details. The harm from such a leak is not necessarily clear at first, but if fraudsters were to access the data of Microsoft customers, they can easily create a malicious scam [9].

Voice data are unlikely to be the first type of data that come to mind with respect to data protection concerns. [10] discuss the concept of “voice ownership” and how this relates to data protection issues. Within their exploration, they illustrate how it is becoming a very real possibility that voice data could be used in a fraudulent way, particularly given the rise in speaker recognition technologies as access mechanisms for digital accounts (such as online banking). To help reduce this type of risk, a group of researchers have launched *The VoicePrivacy Initiative* [11]. This initiative seeks to discover ways to protect the privacy attached to voice data which are used to develop speech technologies. With a specific research event dedicated to it at one of the main annual international speech technology conferences (Interspeech), *The VoicePrivacy Initiative* will be challenging the speech technology community to identify solutions to specific privacy-preserving problems. For example, they may challenge the community to develop voice data anonymisation solutions, or even to invent ways of assessing or measuring how well a system preserves privacy in relation to voice data. While the speech technology community has slightly different demands and objectives, some of the outcomes of *The VoicePrivacy Initiative* may be relevant in the context of forensic speech science.

Underlying the research efforts in [10] and [11] has been the broader increase in awareness across sectors, and the public, with respect to data protection and privacy. This increase in awareness is reflected in legislation and court rulings. The EU General Data Protection Regulation (GDPR, Regulation (EU 2016/679) [12] enacted in UK law under the Data Protection Act (DPA 2018) [13]), in particular, fuelled the attention paid to how and why all kinds of data are stored, used and retained. A central aim of GDPR has been to grant individuals more control over their personal data, and with it has come a greater public awareness of personal data and potential risks attached to having various copies of it in unknown or forgotten places. As a result of GDPR, ordinary users of the internet are now repeatedly asked to consent to their information being stored or used. Organisations can no longer assume consent.

The amount of attention that has been placed on consent can lead to the assumption that this is the only way in which data can be processed lawfully. However, it is recognised by the legislation that consent is not always appropriate, or indeed practical. Article 6 GDPR offers five further options beyond obtaining consent that can enable the lawful processing of data. One of the more relevant options for this paper is labelled "public task", where the processing of the data is necessary for one to perform a task in the public interest or for official functions. It is on this basis that universities can carry out research on data, of course having suitable and secure practical measures in place. On top of this, there are other places within the GDPR that create space for the type of data processing envisaged in this paper. Article 9 GDPR 2(j) states that processing of "special category" data (including biometric data) is allowed when "processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes…". It goes on to emphasise that if data processing is carried out for these purposes, the data must be stored and processed in such a way as to safeguard the fundamental rights and interests of the data subjects.

Additionally, the GDPR and the DPA (2018) recognise that personal data attached to law enforcement require special provisions. Part 3 of the DPA (2018) covers data processing for 'the prevention, investigation, detection or prosecution of criminal offences … including the safeguarding against and the prevention of threats to public security' (Section 31 DPA 2018). Section 35(8)(b) DPA 2018 makes provision for the processing of biometric data for the purpose of uniquely identifying an individual[2]. Biometric data often contributes towards evidence which is used in the Criminal Justice System; when it does, it becomes *forensic data*. The provisions contained within the DPA 2018 and GDPR therefore apply to forensic data. While it might be accepted that, in some ways, forensic data will need to be treated differently to non-forensic data, forensic data is not immune to data protection principles.

[2] We discuss the position of voice data in relation to biometric data in Section 4.1 below.

Indeed, as recognised in Section 42 DPA 2018, extra vigilance and transparency needs to be adopted when dealing with this type of data.

3. Different Perspectives: The balancing act between privacy and improving forensic science

Currently, there is no cohesive regulatory framework that covers the use of biometric and forensic data. Bodies like the Forensic Science Regulator, the Biometrics Commissioner's Office, the Information Commissioner's Office and the Biometrics and Forensics Ethics Group have made some efforts to regulate this area. Although their different perspectives do not neatly combine to form a clear direction, they provide a useful starting point to shape our consideration of voice data protection. We review these different perspectives, in turn, in the following subsections.

3.1 The Biometrics Commissioner

The Biometrics Commissioner is a post that was created to oversee the use and retention of biometric data, with a particular focus on police obtaining, using and retaining DNA samples and fingerprints (rather than focussing on external forensic analysis providers handling these data). This post is filled to satisfy the Protection of Freedoms Act 2012 [14]. Part 1 of the Act deals with the regulation of biometric data, including the destruction, retention and use of such data. The types of biometric data expressly covered are fingerprints and DNA, with an extension to footwear impressions. Footwear impressions do not qualify as "biometric data", but there is still consideration for their protection within the Biometrics Commissioner's role. They therefore provide an interesting reference point when we consider the status of voice as biometric data in Section 4.1 below, and how it should be treated and protected.

Within the Biometrics Commissioner's context, there is emphasis on a *selective and discriminatory approach* to data retention. Within such an approach, all data is filtered to ensure only acceptable material is retained; the retained data is further categorised and allocated a retention period according to specific criteria. For example, if an individual is charged with a “qualifying offence”[3] but not convicted of this offence, their DNA profile and fingerprints may only be retained for three years unless an extension request for a further two years is granted by a District Judge.

The Biometric Commissioner’s Annual Report for 2019 [15] acknowledges the value that storing and using such evidential data for research purposes can bring to public security. Particularly in the context of “new biometrics”, like forensic speech science, retaining case data to form a database can be essential to innovate methods and improve future casework practice. Adopting a discriminatory approach around the retention of data (including the setting of time limits) addresses two purposes simultaneously; it allows for data to be useable for security or research purposes, but also provides the individual concerned with the eventual prospect of clearing personal data from the record.

As we discuss further below, a discriminatory approach that allocates time periods to forensic data samples, based on agreed criteria, seems to be a favourable option among relevant bodies and the public. Given the acceptance of a discriminatory approach in more established forensic disciplines, perhaps this is one that the forensic speech science community could entertain for casework recordings.

3.2 The UK Forensic Science Regulator

As stated in Section 1 of the present paper, the UK Forensic Science Regulator encourages research that advances forensic practice. In line with this, one of the Regulator’s priorities is that scientific analysis procedures should not be static, but should continually improve. In 2016 she produced a protocol [4] that aims to guide how we might use real casework data to help to validate current and new forensic analysis methods. Validation has been communicated as a priority for the forensic science community as it has become more crucial to demonstrate that the methods or techniques that are implemented do indeed achieve what they are claimed to achieve. We can view validation as a specific type of research activity that tests the adequacy of a technique or process for a given purpose.

[3] Qualifying offences are serious offences listed under Section 65A of Police and Criminal Evidence (PACE) Act 1984, ch. 60. There are more than 400 qualifying offences, ranging from murder to kidnapping to offences linked to indecency towards children, etc.

Unlike the Biometric Commissioner's report [15], the *Forensic Science Regulator's Validation Protocol* [4] targets forensic science providers, rather than police forces. Within the protocol, the Forensic Regulator refers to establishing the appropriate processes required to use casework data for validation purposes, but she also refers to establishing the right environment to host these research activities. It states that a clear validation plan should be laid out by the provider and that the provider should seek permission from the Crown Prosecution Service (CPS) or relevant police force to use the case material for validation purposes. In addition, it stipulates that there should be an appropriately qualified individual who is responsible for the protocols and procedures to be followed, as well as for the maintenance of a suitable environment. The forensic science provider also needs a record-keeping system that tracks the storage and use of each specific case data item, the nature and purpose of the validation tasks that they are being included in, and a system that documents how case data is destroyed when it is no longer required. On top of this there is a requirement for the provider to be accredited to ISO 17025/17020 in order for this protocol to apply. The list below provides a summary of the requirements that would be needed to include an instance of casework material in a validation exercise:

1) A validation plan
2) Permission from the CPS or relevant police force
3) A record-keeping procedure for case data storage, the validation activity and destruction details
4) An appropriately qualified individual responsible for the protocols and procedures
5) A suitable environment
6) Accreditation

While validation research is of value, it is not the only type of research that is necessary to progress the forensic speech science field in a meaningful way. For example, it might be that researchers and practitioners wish to explore how one could extract more useful speaker-specific information from a typical "no comment" interview. Currently, when there is a suspect sample from a police interview that largely consists of "no comment", only a limited analysis is generally possible because they provide little coverage of the voice features commonly examined. However, a more in-depth research effort towards these "no comment" recordings may uncover novel aspects of the voice and speech behaviour not currently considered. Opportunity to carry out research on these data is lacking. Validation activities, which could be viewed as a type of demonstrative research, or even a checking exercise, do not allow for the more exploratory research activities that are perhaps warranted in forensic speech science.

3.3 The Information Commissioner

As the UK's independent body that monitors information rights across all kinds of data settings, the Information Commissioner's Office (ICO) provides a lot of valuable information around data protection, and the scope of the ICO extends well beyond this paper's forensic and biometric context. Helpfully, the ICO provides accessible guidance on how to interpret the GDPR in the context of lawful processing of criminal offence data [16].

One particularly pertinent contribution from the ICO is the Information Commissioner's Opinion document [17] that was released in response to the ruling of *R (Bridges) v. Chief Constable of South Wales Police* [18]. Here, a case was brought against South Wales Police in response to their use of live facial recognition technology in a public setting. A number of issues were raised in this case, including:

- whether this was a breach of the right to privacy;
- whether facial data was personal data;
- whether the processing of data was strictly necessary for this purpose
- whether South Wales Police had appropriate documentation in place which covered the processing of sensitive data;
- whether the technology being used was discriminatory.

The court of first instance did not find the use of facial recognition technology to be unlawful. The judges gave a number of reasons including:

- South Wales Police has common law powers to keep peace and prevent crime;
- the technology was deployed in an open and transparent way;
- the data were used for a limited time;
- the technology was used to seek particular individuals (not the Claimant);
- the processing was necessary for the legitimate interests of South Wales Police;
- there was no evidence to suggest that the technology produced discriminatory results.

In her opinion document released in response [17], the ICO suggests that, despite the ruling, there is room for improvement in instances where the police are dealing with sensitive data of this kind. She goes on to raise an interesting point regarding proportionality:

"… the blanket, opportunistic and indiscriminate processing, even for short periods, of biometric data belonging to thousands of individuals in order to identify a few minor suspects or persons of interest is much less likely to meet the high bar contemplated by the [Data Protection Act] 2018. In the Commissioner's Opinion, this is particularly the case if the offences are low level and there may be other less privacy intrusive options available" (p.21).

While it is recognised that the Data Protection Act 2018 caters for law enforcement purposes, the ICO proposes that identifying a small number of suspects at the expense of thousands of individuals' data is not proportionate. This point of proportionality in the context of voice data is further developed below.

The case was appealed and the Court of Appeal [19] overturned the decision arrived at by the court of first instance. The three judges were unanimous in their decision that the technology was used unlawfully by South Wales Police. In giving their reasons, the judges commented on the fact that the conditions of deployment were not clearly defined, and that the technology was not sufficiently tested to identify any inherent biases. The Court of Appeal judgement, no doubt, reflects society's heightened awareness of data protection principles.

3.4 The Biometrics and Forensics Ethics Group

The Biometrics and Forensics Ethics Group (BFEG) is an independent group of experts, sponsored by the UK Home Office, that aims to advise on ethical issues related to biometric and forensic data, and associated technologies. To offer an example of their work, the BFEG set up a working group that is specifically looking into the use of live facial recognition technology, and they have held "evidence gathering days" to make progress in this area [20]. This is in an attempt to investigate all angles of the technology's use and to consider the benefits and dangers of its use. Within their publications (such as [20], [21] and [22]), they echo much of the sentiment that is put forward by the other bodies that have already been covered in this section so far. BFEG highlight the need to respect the privacy of individuals and the need to be open and transparent about the use of data. One theme that emerges among

BFEG's publications, that is not so evident or explicit in the documentation published by the other bodies, is the objective, "to advance justice". It is this theme that resonates with the longer-term objectives of the current paper - that is to advance practice in forensic speech analysis.

3.5 The Public

It is also crucial for the public to be taken into account when considering both sides of the current topic: data protection and improving forensic science. There have been some public attitudes studies that have aimed to capture public opinion on such matters.

In [23], one hundred informants in New Zealand took part in a survey that questioned their knowledge and attitudes towards having a DNA database for forensic purposes. Generally speaking, the participants recognised the potential of such a database as a "crime-fighting tool", but a large proportion of the participants still expressed concern about its use. In particular, 60% of the participants were concerned that DNA might be used for another purpose, and 59% were concerned about mistakes being made (e.g. false identifications).

Another example of a public attitudes survey was initiated by the Ada Lovelace Institute who published findings of a survey distributed to over 4000 informants that targeted the use of live facial recognition technology [24]. The survey revealed public concerns for normalised use of surveillance technologies, but it also revealed that the majority of respondents supported the use of such technology for police criminal investigations as the public can generally see the security benefits.

While these surveys may capture a snapshot of public attitudes towards the topic, they are not designed to capture the depth that is perhaps required for such a complex issue. The Ada Lovelace Institute recently adopted a more in-depth process for capturing public attitudes by establishing the Citizens' Biometrics Council, which consisted of 50 members of the public. The Council participated in numerous workshops and consultations with experts, allowing the Council to meaningfully debate issues around biometric technologies, in particular. This comprehensive process led to a report that contains a set of resulting recommendations [25]. A key theme that transpires from the recommendations is the lack of current legislation and regulation with regards to biometric technologies. The Council calls for developments in this area (also, see Footnote 1).

A pair of recent rulings that are relevant to the present discussion around the public's perspective are that of *Gaughran v. Chief Constable of the Police Service of Northern Ireland* [26] and *Gaughran v. the United Kingdom* [27]. The case involves Mr Gaughran who was arrested in October 2008 for drink driving. After a positive breath sample, his photograph was taken alongside a DNA sample and fingerprints. Mr Gaughran pleaded guilty and his conviction was spent in 2013. In 2015, Mr Gaughran challenged the indefinite retention of his personal data, on the basis that it was disproportionate and a breach of the right to private and family life. The Supreme Court found that the indefinite retention of his data was a breach of his right to privacy; however, the breach was held to be proportionate [26]. In contrast, the European Court of Human Rights (ECHR) ruled that the breach was disproportionate [27]. One of the reasons given by the ECHR was that the availability of new technology means that these data can be used for new, previously unforeseen purposes (e.g. the use of photographs in facial recognition software). The implications of data retention in 2008 are not the same as the implications of data retention in 2020, therefore altering what might be considered to be proportionate through time.

Against the backdrop of the *Gaughran* rulings, [28] share findings of a public attitudes survey that asked 201 people for their views on retaining DNA profiles of convicted individuals. Their overall conclusion suggests that people would be accepting of a "discriminatory" regime that draws a distinction between individuals who were convicted of serious offences and less serious offences. 83% of the respondents were supportive of long-term retention of DNA profiles in cases where a serious offence had been committed, whereas 47% of the respondents supported long-term retention where a more minor offence had been committed. Likewise, the responses reported in [23] show similar support for a discriminatory approach, this time distinguishing between conviction and arrest. To exemplify, 89% of the participants were in favour of a DNA database for individuals convicted of a violent crime, while 44% of the participants supported the idea of a DNA database for individuals who are suspected of a crime. The type of discriminatory approach outlined in the Biometrics Commissioner's Annual Report [15] appears to resonate with the trends emerging from these public attitudes surveys.

3.6 An Overview of Perspectives

All of the perspectives and emerging themes addressed in Sections 3.1 – 3.4 are relevant to developing a way forward in the context of forensic voice data. Table 1 provides a summary overview:

Table 1: Summary of the key points that have emerged from a review of the relevant bodies and documents.

Relevant Body	Priorities or focus	Document(s) referred to	Comments on relevant points in the document(s)
Biometrics Commissioner	DNA and fingerprints used by police forces	Biometrics Commissioner Annual Report for 2019 [15]	• The adoption of a discriminatory approach to retaining DNA and fingerprint evidence on record.
UK Forensic Regulator	Research to continually improve practice and capability	Forensic Science Regulator Protocol: Validation – Use of Casework Material, FSR-P-300 [4]	• Presents practical guidance on how to legitimately store real forensic data. • The guidance is quite broad to allow for its application to many forensic disciplines. • Targets validation research only, which does not account for more exploratory research.
Information Commissioner	Oversees general data and information rights matters	Published Opinion in response to the *R (Bridges) v Chief Constable of South Wales Police* ruling [17]	• Proposes that the collecting and retaining of thousands of people's data for the sake of identifying a small number of minor suspects is disproportionate. • Also draws attention to the ethics of opportunism in retaining data.
Biometrics and Forensics Ethics Group	Independent group of experts that aims to advise on ethical issues related to biometric and forensic data and associated technologies	Biometrics and Forensics Ethics Group Annual Reports and their Ethical Principles Document [20, 21, 22]	• Echoes the points raised by other bodies regarding the challenges of weighing up the privacy rights of individuals against the benefits of public security in relation to retaining forensic and biometric data. • There is a stronger focus on the longer-term benefit of "advancing justice" that

			may be brought about by retaining forensic and biometric data.
The public	NA	• The Citizen's Biometrics Council Report [25] • Amankwaa and McCartney (2020) – reporting findings of a public attitudes survey [28]	• Overall suggest that a discriminatory approach to forensic or biometric data would be largely acceptable to the public. • Many respondents believed that longer-term retention of DNA profiles is acceptable when the individual has been convicted of a serious offence.

4. The case of forensic voice data

The key considerations in relation to data protection principles and forensic data that have emerged from Section 3 are:

- the need to go beyond validation research to carry out more exploratory work
- "opportunism" in data retention
- the consideration of proportionality in forensic data retention
- the implementation of a discriminatory approach to data retention

This section takes the above considerations and points out the specific challenges and issues that arise when dealing with forensic voice data, starting with a consideration of whether voice is biometric data or not in Section 4.1. Section 4.2 puts forward a discriminatory approach to the storage and retention of voice data, as well as the issues involved. Section 4.3 outlines practical steps that could be followed to make it more acceptable to use forensic voice recordings for research purposes.

4.1 Voice as a biometric?

Much of the relevant literature, documentation and guidance applies to "biometric" data. It is therefore important to consider whether voice falls within this data category or not. The definition of biometric data that is provided within the GDPR is as follows:

"personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person" (Article 4(14)).

According to this definition, voice data does not strictly apply. While voice can provide useful information with regards to an individual's identity, it cannot go so far as to "uniquely identify" an individual.

[29] comprehensively discuss the possible ambiguity of "unique identification" in relation to voice data. A literal interpretation of this phrase assumes the highest "threshold of identification" (i.e. identifying an individual to the exclusion of all others). However, this would not be an appropriate reading in the context of voice. Forensic speech analysis does not achieve the same strength of evidence that can be achieved with, say, DNA analysis. We therefore cannot comfortably place voice data in the biometric category. While this could easily be seen as a subtle distinction and a minor point, it is an extremely important one for the current discussion. An overestimation of the potential for voice to uniquely identify an individual could unnecessarily prevent the use of forensic voice recordings for valuable research.

Having said this, it would be wrong to suggest that voice does not resonate with the definition of biometric data at all. There are still links between an individual's voice and their identity. It is perhaps more appropriate to think about "biometrics" on a sliding scale, rather than to adopt a 'black or white' type of categorisation. In the Biometrics Commissioner's 2019 report [15], a distinction is drawn between the likes of DNA and footwear impressions. It is acknowledged that footwear impressions are not a biometric. Footwear impressions cannot "uniquely identify" an individual, and therefore a database of footwear impressions could not act as a database for "matching" in the same way as DNA does in the National DNA Database (NDNAD) database. In view of this, the law around the retention of footwear impressions is less specific, stipulating that 'Impressions of footwear may be retained for as long as is necessary for purposes related to the prevention or detection of crime, the investigation of an offence or the conduct of a prosecution.' (Part 1 Section 15 of [14]). That said, the fact that non-biometric data is included in Protection of Freedoms Act indicates that the concept of data minimisation (i.e. the fact that data should not be retained for longer than necessary) is not only relevant to biometric data, but also non-biometric data, and indeed all data which falls in between, i.e. voice data.

It is worth noting that "voice data" can encapsulate many different types of data. There are the actual voice samples themselves contained within audio recordings. However, voice data also include the voice representations generated by automatic speaker recognition systems, and the voice profiles arrived at by the human analyst (as represented in practitioners' analysis notes). Similar data type distinctions are seen with respect to DNA, where there is the physical DNA sample, as well as the DNA profile. The distinction between DNA profiles and samples lies in DNA profiles consisting of strings of numbers and letters that can be meaningfully compared against other DNA profiles in order to make matches. DNA samples, on the other hand, contain biological and genetic material. The Protection of Freedoms Act 2012 differentiates between physical DNA samples and DNA profiles, with samples being deleted within six months of being taken[4], while profiles are obtained and stored on the National DNA Database (NDNAD). The same sample-profile distinction cannot be made with regards to voice data, as the voice profiles do not even come close to DNA profiles with respect to their power to identify an individual. As this same distinction cannot be made, it would be disproportionate to adopt the DNA data retention framework to voice (neither profiles nor samples). Furthermore, any data protection framework that is put in place for voice needs to apply to voice samples as it is the voice samples that would enable the type of research that can lead to meaningful developments within forensic speech science. Given their very limited potential to identify an individual, voice profiles are less of a data protection concern.

4.2 Proposing a discriminatory approach to retaining voice data

The discussion in Section 4.1 leads to another theme that emerged from the Biometrics Commissioner's report, and that is the use, by police forces, of a discriminatory approach to retaining biometric data. This theme also emerged from the public attitudes surveys and the Gaughran case discussed above. It links to the issue of proportionality, whereby it may be seen as unnecessary to retain data samples from individuals in instances of "more minor

[4] With the exception of the DNA sample forming part of evidence in court, under Criminal Procedure and Investigations Act 1996 [30].

cases”. A discriminatory approach is adopted for DNA and fingerprinting, and the decision around how long these data are retained depends on the nature of the offence and whether the person has been convicted. As the Gaughran case has revealed, achieving the “right” retention periods and guidelines for different case categories is not necessarily straightforward. The Biometrics and Forensics Ethics Group (BFEG) acknowledge that more work needs to be carried out on the topic of data retention periods [22].

In principle, it is possible to implement a discriminatory approach to retaining forensic voice data. It is feasible to destroy speech recordings after given durations, and to develop a data review system to assist with this. Indeed, there is Home Office guidance that puts forward retention periods of material seized for forensic examination [31]. However, because of the nature of speech material and the channels through which forensic speech analysis is carried out, this guidance becomes challenging to implement. The current arrangement for the provision of forensic voice analysis is that there is a reliance on private providers who get contracted work by the police. It is likely to be these providers that form research databases of forensic voice data. Unlike the police, private providers do not necessarily receive information in relation to the offence; nor do they routinely find out whether a person was convicted, acquitted or indeed charged. It is this information that would be required if we were to implement a discriminatory approach in this area.

There is another key consideration to take into account in relation to voice data, which further complicates matters: so-called “secondary subjects”. This is a consideration that is raised among the BFEG’s *Ethical Principles* [21]. The BFEG provides the example that family members of the individuals whose data is retained may also be at risk or affected in some way. In the case of forensic voice data, there are two types of secondary subject data to account for. The first type relates to voice recordings of secondary subjects. The recordings that a forensic speech analyst receives regularly contain voices of multiple speakers (not just the speaker of interest), and it follows that these voices would require protection. This could simply be achieved by not retaining speech from secondary subjects. Or, if it were the case that the secondary subject’s speech had to be retained, it could be artificially disguised (using voice conversation technology, for example). The second type of secondary subject data is that the voice evidence itself might hold further information about other individuals beyond the primary person of interest. Police interview recordings, as well as recordings relating to an offence, can contain comprehensive information about an event or about other people.

That information might relate to personal information such as names, dates-of-birth and addresses which directly point to individuals. However, there is also indirect personal information in that seemingly neutral aspects of the spoken content can nevertheless point towards an individual (e.g. a party happening at a specific pub at a specific time). Certain listeners, with the necessary knowledge, may be able to guess whether an individual is being described.

While we recognise the importance of protecting secondary subjects, the safeguarding need not be turned into an indomitable barrier. We should accept that we can never completely eliminate the risk that an individual is going to be traceable by the contextual information contained within a recording; it is about finding a pragmatic solution to minimise the risk. A feasible solution would be to form a set of anonymisation and redaction criteria. These criteria might state that factual information such as name, date-of-birth and address should generally be redacted. With respect to the indirect information, whether this needs to be redacted could be considered on a case-by-case basis.

For a discriminatory approach to work in forensic speech science, we would need to develop a smooth communication channel between the police force and the provider to ensure that the relevant information is communicated between the two parties. We would also need to develop redaction and anonymisation criteria. These measures would be included in a set of data protection policies. Having these measures in place could absorb some of the concern around the retention and use of these data, thereby making their use for research more acceptable. Section 4.3 below continues to outline the set of measures a provider could implement.

4.3 Applying existing recommendations to voice data

As explained in Section 3.2, the *Forensic Science Regulator Validation Protocol* [4] provides welcome practical guidance to allow for validation trials to be carried out on casework material. The Protocol provides a useful starting point to move us beyond validation research and facilitate more exploratory research activities. We have broken down the guidance into four main areas and elaborated on how we suggest each could be applied to forensic voice data:

1) *Creating a trustworthy and responsible environment*

There are a number of components to creating a trustworthy and responsible environment:

- Firstly, given that voice data is processed and analysed digitally, having robust cybersecurity measures in place is key. Following the recommendations of schemes like the government-backed Cyber Essentials[5] package can meet required cybersecurity standards.
- Secondly, having an appropriately qualified individual who is responsible and accountable for the data security of the organisation will also contribute to the right environment. As part of this role, the individual will oversee the implementation of anonymisation criteria. The sensitive and confidential information should not go beyond the responsible individual.
- Finally, a commitment to transparency and openness will also be key to creating a trustworthy environment. This can be achieved by creating accessible research plans that clearly state the purpose(s) of the data retention. By specifying the research purposes, and sticking to them, only data that is needed for those purposes will be stored, thereby observing the principle of *data minimisation*. This simultaneously avoids the Information Commissioner's concerns around "opportunistic" data collection and storage, which refers to more aimless and vague (but not necessarily bad) intentions for the data.

2) *Comprehensive documentation processes*

Details about how and when data is stored, used and destroyed should be documented. This information could be within an organisation's data protection policy (which includes details about the IT security), in the data research plan, or in the organisation's Standard Operating Procedures.

3) *Gaining permission*

An agreement should be reached between relevant parties. Ideally, a *Data Sharing Policies* agreement would be put in place that clearly outlines the specific uses and users of the data. The agreement would serve the purpose of both obtaining

[5] https://www.ncsc.gov.uk/cyberessentials/overview [accessed: 06/04/2021]

permission to use the data, as well as explaining the nature of the planned research activities. In some cases, there are existing agreements between the forensic services provider and an instructing police force, where it is stipulated that the data should only be used to fulfil the service (i.e. the forensic analysis). In these cases, it should be explored whether permission can be gained to use the data for another purpose, and the agreements amended accordingly. In instances where the forensic provider works with an academic institution, similar agreements should be put in place.

4) Accreditation

For the FSR Validation Protocol to apply to forensic providers, it stipulates that providers should be accredited. This hugely limits the number of forensic providers who could engage with validation activities, never mind more exploratory research activities (particularly in the "niche" forensic disciplines). In the UK, at least, it may well be the case that forensic providers are taking steps towards accreditation for certain aspects of their work, but this is still very much an ongoing effort. This does not mean that providers are not following responsible procedures and protocols. An absence of a 'stamp of approval' by way of official accreditation to ISO 17025/17020 should not be taken to indicate that providers are not ensuring that their practices are to standard. There is also a cyclical aspect to this as it is part of the accreditation process for the provider to demonstrate active engagement with their field and to push for progress within it. It would therefore seem counterintuitive for an absence of accreditation to be a block on engaging with casework-relevant research, especially if a provider has appropriate practices and conditions that can aid the progress of the field.

5. Discussion and Conclusion

At the very least, this paper has opened up the conversation around data protection issues with a specific focus on forensic voice evidence. Themes that have been prominent in this navigation are: the definition of "biometric data", proportionality, a discriminatory approach to data retention and practical solutions to using casework data for research.

While “validation research” has its place, there are great benefits to be drawn from carrying out more exploratory and innovative research. This could appear “opportunistic”. It is clear, however, that the intentions behind the present paper align with a direction encouraged by the UK Forensic Science Regulator, which is to continue research efforts in order to improve the quality of forensic science provisions. This will in turn “advance justice”. We propose that it is possible to carry out research that is more exploratory in nature while at the same time adhering to data protection principles. We have suggested practical solutions in this regard, such as creating the right environment for forensic voice data retention and developing clear data research plans. Taking public attitudes research and existing frameworks into account, it seems that a discriminatory approach to retaining forensic voice data is likely to be the most amenable. We are keen to continue discussions on what a discriminatory data retention approach could look like in forensic speech science.

The purpose of the current work has been to carve out solutions to access forensic voice data for research activities, but making forensic voice data available would be of benefit elsewhere. Bringing real casework data into teaching and training contexts is an obvious application. Forensic speech science is now taught by a small number of higher education institutions at both undergraduate and postgraduate level. Graduates of these courses and modules have been recruited into forensic speech analysis roles for private providers and also in the public sector. It is highly desirable that students on these modules and courses are taught using real casework data in order to better-prepare them for potential discipline-specific opportunities. There are additional factors to keep in mind when considering real casework recordings for this purpose (for example, it would involve exposing these data to a larger audience rather than keeping them within a very small research team). However, pursuing the integration of casework data into teaching would be in the interests of the field and those who benefit from the field.

Finally, the current paper exists as a result of there not being a single port-of-call to ask for advice or find clear guidance in relation to using forensic voice data for research and development purposes. Ideally, there would be a single “go-to” authority that oversees the types of data matters discussed here and it is hoped that an authority will be identified or established in the near future. In the meantime, we are confident that a comprehensive demonstration of data protection measures and a clear move towards openness and transparency could achieve a satisfactory balance between data protection principles and research developments.

References

[1] Nolan, F., McDougall, K., de Jong, G., and Hudson, T. (2009). The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *Forensic Linguistics*, 16 (1).

[2] Watt, D., Harrison, P., Hughes, V., French, J. P., Llamas, C., Braun, A. and Robertson, D. (2020) Assessing the effects of accent-mismatched reference databases on the performance of an automatic speaker recognition system. *International Journal of Speech, Language and the Law*. 27. 1-34.

[3] Gold, E., Ross. S., and Earnshaw, K. (2018) The 'West Yorkshire Regional English Database': Investigations into the generalizability of reference populations for forensic speaker comparison casework. In *Proceedings of Interspeech 2018: September 2-6 2018, Hyderabad* (pp. 2748-2752).

[4] Forensic Science Regulator Protocol: Validation – Use of Casework Material, FSR-P-300, Issue 2, The Forensic Science Regulator, Birmingham, UK.

[5] Roberts, L. (2012). A forensic phonetic study of the vocal responses of individuals in distress. PhD thesis. University of York, UK.

[6] *Human Rights Act 1998,* ch. 42, Schedule 1, Part 1, Article 8. URL: https://www.legislation.gov.uk/ukpga/1998/42/schedule/1/part/I/chapter/7. Accessed: 12/11/2020.

[7] National Cyber Security Centre. (2020). NCSC statement: EasyJet cyber incident. URL: https://www.ncsc.gov.uk/news/easyjet-incident. Accessed: 9/11/2020.

[8] Bischoff, P. (2020). Report: 250 million Microsoft customer service and support records exposed on the web. URL: https://www.comparitech.com/blog/information-security/microsoft-customer-service-data-leak/. Accessed: 16/10/2020.

[9] Winder, D. (2020). Microsoft Security Shocker as 250 million customer records exposed online. Forbes. URL: https://www.forbes.com/sites/daveywinder/2020/01/22/microsoft-security-shocker-as-250-million-customer-records-exposed-online/#91076484d1b3. Accessed: 16/10/2020.

[10] Watt, D., Harrison, P. and Cabot-King, L. (2019). Who owns your voice? Linguistic and legal perspectives on the relationship between vocal distinctiveness and the rights of the individual speaker. *International Journal of Speech, Language and the Law,* 26 (2), 137-180.

[11] Tomaschenko, N., Srivastava, B. M. L., Wang, X., Vincent, E., Nautsch, A., Yamagishi, J., Evans, N., Patino, J., Bonastre, J-F, Noe, P-G, Todisco, M. (2020) Introducing the VoicePrivacy Initiative, *Interspeech.* Shanghai, China.

[12] General Data Protection Regulation. European Parliament and Council of European Union (2016) *Regulation (EU) 2016/679*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN. Accessed: 09/11/2020.

[13] *Data Protection Act 2018*, ch. 12. URL: http://www.legislation.gov.uk/ukpga/2018/12/contents/enacted. Accessed: 09/11/2020.

[14] *Protection of Freedoms Act 2012*, ch. 9. URL: https://www.legislation.gov.uk/ukpga/2012/9/contents/enacted. Accessed: 12/11/2020.

[15] Wiles, P. (2020) Annual Report 2019: Commissioner for the Retention and Use of Biometric Material. Office of the Biometrics Commissioner.

[16] Information Commissioner's Office (2018). Guide to the General Data Protection Regulation. URL: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-

general-data-protection-regulation-gdpr/lawful-basis-for-processing/criminal-offence-data/ Accessed: 30/10/2020.

[17] Information Commissioner's Office (2019). Information Commissioner's Opinion: The use of live facial recognition technology by law enforcement in public places. URL: https://ico.org.uk/media/about-the-ico/documents/2616184/live-frt-law-enforcement-opinion-20191031.pdf

[18] R (Bridges) v. Chief Constable of South Wales Police [2019] EWHC 2341 (Admin).

[19] R (Bridges) v Chief Constable of South Wales Police [2020] EWCA Civ 1058.

[20] Biometrics and Forensics Ethics Group. (2021). Annual Report for 2019-2020 URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/964471/Biometrics_and_Forensics_Ethics_Group_annual_report_2019_-_2020.pdf Accessed: 06/04/2021.

[21] Biometrics and Forensics Ethics Group (2020a). Ethical Principles Document. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/946996/BFEG_Principles_Update_December_2020.pdf Accessed: 06/04/2021.

[22] Biometrics and Forensics Ethics Group. (2020b). Annual Report for 2018 URL: https://www.gov.uk/government/publications/biometrics-and-forensics-ethics-group-annual-report-2018 Accessed 12/11/2020.

[23] Curtis. C. (2009). Public Perceptions and Expectations of the Forensic Use of DNA: Results of a preliminary study. *Bulletin of Science, Technology and Society*. 29. 313-324.

[24] Ada Lovelace Institute (2019). Beyond face value: public attitudes to facial recognition technology [PDF],
available from: https://www.adalovelaceinstitute.org/wp-content/uploads/2019/09/Public-attitudes-to-facial-recognition-technology_v.FINAL_.pdf. Accessed 12/11/2020.

[25] The Citizens' Biometrics Council (2021). Recommendations and findings of a public deliberation on biometrics technology, policy and governance. The Ada Lovelace Institute. URL: https://www.adalovelaceinstitute.org/project/citizens-biometrics-council/

[26] Gaughran v. Chief Constable of the Police Service of Northern Ireland [2015] UKSC 29.

[27] Gaughran v the United Kingdom [2020] ECHR 144.

[28] Amankwaa, A. and McCartney, C. (2020). Gaughran vs the UK and public acceptability of forensic biometrics retention. *Science and Justice*, 60 (3): 204-205.

[29] 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 Proceedings of Interspeech, September 15-19 2019, Graz, Austria.

[30] *Criminal Procedure and Investigations Act 1996*. URL: https://www.legislation.gov.uk/ukpga/1996/25/contents

[31] UK Home Office. (2019). Storage, retention and destruction of records and materials seized for forensic examination. URL: https://www.gov.uk/government/publications/storage-retention-and-destruction-of-records-and-materials-seized-for-forensic-examination/storage-retention-and-destruction-of-records-and-materials-seized-for-forensic-examination-accessible-version. Accessed: 30/10/2020.