Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference

Jacob Leon Kröger, Otto Hans-Martin Lutz, Philip Raschke

Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference

Jacob Leon Kröger[1,2], Otto Hans-Martin Lutz[1,2,3], Philip Raschke[1]

[1] Technische Universität Berlin
Straße des 17. Juni 135, 10623 Berlin, Germany
{kroeger,philip.raschke}@tu-berlin.de

[2] Weizenbaum Institute for the Networked Society, Berlin, Germany
[3] Fraunhofer Institute for Open Communication Systems, Berlin, Germany

Abstract. Internet-connected devices, such as smartphones, smartwatches, and laptops, have become ubiquitous in modern life, reaching ever deeper into our private spheres. Among the sensors most commonly found in such devices are microphones. While various privacy concerns related to microphone-equipped devices have been raised and thoroughly discussed, the threat of unexpected inferences from audio data remains largely overlooked. Drawing from literature of diverse disciplines, this paper presents an overview of sensitive pieces of information that can, with the help of advanced data analysis methods, be derived from human speech and other acoustic elements in recorded audio. In addition to the linguistic content of speech, a speaker's voice characteristics and manner of expression may implicitly contain a rich array of personal information, including cues to a speaker's biometric identity, personality, physical traits, geographical origin, emotions, level of intoxication and sleepiness, age, gender, and health condition. Even a person's socioeconomic status can be reflected in certain speech patterns. The findings compiled in this paper demonstrate that recent advances in voice and speech processing induce a new generation of privacy threats.

Keywords: Audio, Voice, Speech, Microphone, Privacy, Inference, Side channel

1 Introduction

Since the invention of the phonograph in the late 19th century, it has been technically possible to record and reproduce sounds. For a long time, this technology was exclusively used to capture pieces of audio, such as songs, audio tracks for movies, or voice memos, and for the telecommunication between humans. With recent advances in automatic speech recognition, it has also become possible and increasingly popular to interact via voice with computer systems [96].

Microphones are ubiquitous in modern life. They are present in a variety of electronic devices, including not only phones, headsets, intercoms, tablet computers, dictation machines and baby monitors, but also toys, household appliances, laptops, cameras, smartwatches, cars, remote controls, and smart speakers.

There is no question that microphone-equipped devices are useful and important in many areas. It is hard to imagine a future, or even a present, without them. However, as a growing proportion of audio recordings is disseminated through insecure communication networks and processed on remote servers out of the user's control, the ubiquity of microphones may pose a serious threat to consumer privacy. Research and public debates have addressed this concern, with published reports looking into technical and legal aspects regarding data collection, processing, and storage, as well as access and deletion rights of the data subjects [18, 32, 96]. Yet, the recent privacy discourse has paid too little attention to the wealth of information that may unexpectedly be contained in audio recordings.

Certain characteristics of human speech can carry more information than the words themselves [94]. With the help of intelligent analysis methods, insights can not only be derived from a speaker's accent, dialect, sociolect, lexical diversity, patterns of word use, speaking rate and rhythms, but also from acoustic properties of speech, such as intonation, pitch, perturbation, loudness, and formant frequencies. A range of statistics can be applied to extract hundreds or even thousands of utilizable speech parameters from just a short sequence of recorded audio [19, 80].

Based on literature of diverse scientific disciplines, including signal processing, psychology, neuroscience, affective computing, computational paralinguistics, speech communication science, phonetics, and biomedical engineering, section 2 of this paper presents an overview of sensitive inferences that can be drawn from linguistic and acoustic patterns in audio data. Specifically, we cover inferences about a user's biometric identity (section 2.1), body measures (sect. 2.2), moods and emotions (sect. 2.3), age and gender (sect. 2.4), personality traits (sect. 2.5), intention to deceive (sect. 2.6), sleepiness and intoxication (sect. 2.7), native language (sect. 2.8), physical health (sect. 2.9), mental health (sect. 2.10), impression made on other people (sect. 2.11), and socioeconomic status (sect. 2.12). Additionally, we examine information that can be extracted from the ambient noise and background sounds in a voice recording (sect. 2.13). Section 3 provides a discussion of the presented findings with regard to their limitations and societal implications, followed by a conclusion in section 4.

2 Inference of Personal Information from Voice Recordings

Based on experimental studies from the academic literature, this section presents existing approaches to infer information about recorded speakers and their context from speech, non-verbal human sounds, and environmental background sounds commonly found in audio recordings. Where available, published patents are also referenced to illustrate the current state of the art and point to potential real-world applications.

Fig. 1 provides an introductory overview of the types of audio features and the categories of inferences discussed in this paper.

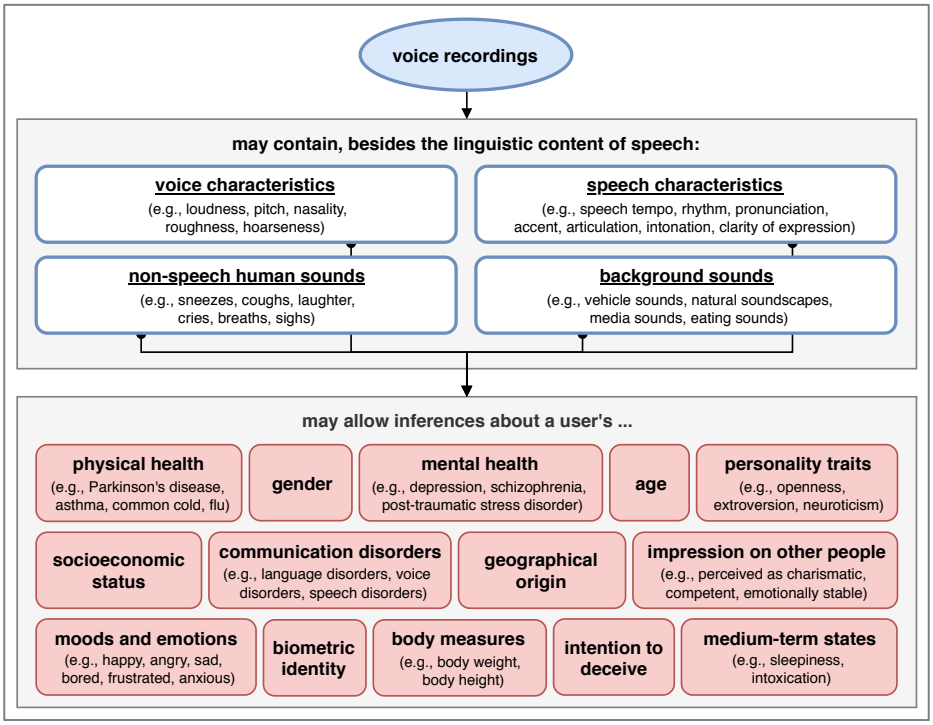

Fig. 1. Overview of some sensitive attributes discernable from speech data.

2.1 Speaker Recognition

Human voices are considered to be unique, like handwriting or fingerprints [100], allowing for the biometric identification of speakers from recorded speech [66]. This has been shown to be possible with speech recorded from a distance [71] and with multi-speaker recordings, even under adverse acoustic conditions (e.g., background noise, reverb) [66]. Voice recognition software has already been transferred into patents [50] and is being applied in practice, for example to verify the identity of telephone customers [40] or to recognize users of virtual assistants like Amazon Alexa [1].

Mirroring the privacy implications of facial recognition, voice fingerprinting could be used to automatically link the content and context of sound-containing media files to the identity of speakers for various tracking and profiling purposes.

2.2 Inference of Body Measures

Research has shown that human listeners can draw inferences about body characteristics of a speaker based solely on hearing the target's voice [42, 55, 69]. In [42], voice-based estimates of waist-to-hip ratio (WHR) of female speakers predicted the speaker's actual WHR, the estimated shoulder-to-hip ratio (SHR) of male speakers predicted the speaker's actual SHR measurements. In another study, human evaluators estimated the

body height and weight of strangers from a voice recording almost as well as they did from a photograph [55].

Various attempts have been made to identify the acoustic voice features that enable such inferences [25, 29, 69]. In women, relationships were discovered between voice parameters, such as subharmonics and frequency pertubation, and body features, including weight, height, body mass index, and body surface area [29]. Among men, individuals with larger body shape, particularly upper body musculature, are more likely to have low-pitched voices, and the degree of formant dispersion in male voices was found to correlate with body size (height and weight) and body shape (e.g., waist, chest, neck, and shoulder circumference) [25].

Although research on the speech-based assessment of body configuration is not as advanced as other inference methods covered in this paper, corresponding algorithms have already been developed. For instance, researchers were able to automatically estimate the body height of speakers based on voice features with an accuracy of 5.3 centimeters, surpassing human performance at this task [69].

Many people feel uncomfortable sharing their body measurements with strangers [12]. The researchers who developed the aforementioned approach for speech-based body height estimation suggest that their algorithm could be used for "applications related to automatic surveillance and profiling" [69], thereby highlighting just some of the privacy threats that may arise from such inference possibilities.

2.3 Mood and Emotion Recognition

There has been extensive research on the automatic identification of emotions from speech signals [21, 23, 53, 95, 99]. Even slight changes in a speaker's mental state invoke physiological reactions, such as changes in the nervous system or changes in respiration and muscle tension, which in turn affect the voice production process [20]. Besides voice variations, it is possible to automatically detect non-speech sounds associated with certain emotional states, such as crying, laughing, and sighing [4, 23].

Some of the moods and emotions that can be recognized in voice recordings using computerized methods are happiness, anger, sadness, and neutrality [86], sincerity [37], stress [95], amusement, enthusiasm, friendliness, frustration, and impatience [35], compassion and sarcasm [53], boredom, anxiousness, serenity, and astonishment [99]. By analyzing recorded conversations, algorithms can also detect if there is an argument [23] or an awkward, assertive, friendly, or flirtatious mood [82] between speakers.

Automatic emotion recognition from speech can function under realistic noisy conditions [23, 95] as well as across different languages [21] and has long been delivering results that exceed human performance [53]. Audio-based affect sensing methods have already been patented [47, 77] and translated into commercial products, such as the voice analytics app *Moodies* [54].

Information about a person's emotional state can be valuable and highly sensitive. For instance, Facebook's ability to automatically track emotions was a necessary precondition for the company's 2014 scandalous experiment in which the company observed and systematically manipulated mental states of over 600,000 users for opaque purposes [14].

2.4 Inference of Age and Gender

Numerous attempts have been made to uncover links between speech parameters and speaker demographics [26, 34, 48, 92]. A person's gender, for instance, can be reflected in voice onset time, articulation, and duration of vowels, which is due to various reasons, including differences in vocal fold anatomy, vocal tract dimensions, hormone levels, and sociophonetic factors [92]. It has also been shown that male and female speakers differ measurably in word use [26]. Like humans, computer algorithms can identify the sex of a speaker from a voice sample with high accuracy [48]. Precise classification results are achieved even under adverse conditions, such as loud background noise or emotional and intoxicated speech [34].

Just as the gender of humans is reflected in their anatomy, changes in the speech apparatus also occur with the aging process. During puberty, vocal cords are thickened and elongated, the larynx descends, and the vocal tract is lengthened [15]. In adults, age-related physiological changes continue to systematically transform speech parameters, such as pitch, formant frequencies, speech rate, and sound pressure [28, 84].

Automated approaches have been proposed to predict a target's age range (e.g., child, adolescent, adult, senior) or actual year of birth based on such measures [28, 85]. In [85], researchers were able to estimate the age of male and female speakers with a mean absolute error of 4.7 years. Underlining the potential sensitivity of such inferred demographic information, unfair treatment based on age and sex are both among the most prevalent forms of discrimination [24].

2.5 Inference of Personality Traits

Abundant research has shown that it is possible to automatically assess a speaker's character traits from recorded speech [3, 79, 80, 88]. Some of the markers commonly applied for this purpose are prosodic features, such as speaking rate, pitch, energy, and formants [68] and characteristics of linguistic expression [88].

Existing approaches mostly aim to evaluate speakers along the so-called "Big Five" personality traits (also referred to as the "OCEAN model"), comprising openness, conscientiousness, extroversion, agreeableness, and neuroticism [88]. The speech-based recognition of personality traits is possible both in binary form (high vs. low) and in the form of numerical scores [79]. High estimation accuracies have been achieved for all OCEAN traits [3, 80, 88].

Besides the Big Five, voice and word use parameters have been correlated with various other personality traits, such as gestural expressiveness, interpersonal awkwardness, fearfulness, and emotionality [26]. Even culture-specific attributes, such as the extent to which a speaker accepts authority and unequal power distribution, can be inferred from speech data [101].

It is well known that personality traits represent valuable information for customer profiling in various industries, including targeted advertising, insurance, and credit risk assessment – with potentially harmful effects for the data subjects [17, 18]. Some data analytics firms also offer tools to automatically rate job applicants and predict their likely performance based on vocal characteristics [18].

2.6 Deception Detection

Research has shown that the veracity of verbal statements can be assessed automatically [60, 107]. Among other speech cues, acoustic-prosodic features (e.g., formant frequencies, speech intensity) and lexical features (e.g., verb tense, use of negative emotion words) were found to be predictive of deceptive utterances [67]. Increased changes in speech parameters were observed when speakers are highly motivated to deceive [98].

Speech-based lie detection methods have become effective, surpassing human performance [60] and almost reaching the accuracy of methods based on brain activity monitoring [107]. There is potential to further improve the classification performance by incorporating information on the speaker's personality [2], some of which can be inferred from voice recordings as well (as we have discussed in section 2.5).

The growing possibilities of deception detection may threaten a recorded speaker's ability to use lies as a means of sharing information selectively, which is considered to be a core aspect of privacy [63].

2.7 Detection of Sleepiness and Intoxication

Medium-term states that affect cognitive and physical performance, such as fatigue and intoxication, can have a measurable effect on a speaker's voice. Approaches exist to automatically detect sleepiness from speech [19, 89]. There is even evidence that certain speech cues, such as speech onset time, speaking rate, and vocal tract coordination, can be used as biomarkers for the separate assessment of cognitive fatigue [93] and physical fatigue [19].

Similar to sleepiness and fatigue, intoxication can also have various physiological effects, such as dehydration, changes in the elasticity of muscles, and reduced control over the vocal apparatus, leading to changes in speech parameters like pitch, jitter, shimmer, speech rate, speech energy, nasality, and clarity of pronunciation [5, 13]. Slurred speech is regarded as a hallmark effect of excessive alcohol consumption [19].

Based on such symptoms, intoxicated speech can be automatically detected with high accuracy [89]. For several years now, systems have been achieving results that are on par with human performance [13]. Besides alcohol, the consumption of other drugs such as ±3,4-methylenedioxymethamphetamine ("MDMA") can also be detected based on speech cues [7].

2.8 Accent Recognition

During childhood and adolescence, humans develop a characteristic speaking style which encompasses articulation, phoneme production, tongue movement, and other vocal tract phenomena and is mostly determined by a person's regional and social background [64]. Numerous approaches exist to automatically detect the geographical origin or first language of speakers based on their manner of pronunciation ("accent") [9, 45, 64].

Research has been done for discriminating accents within one language, such as regional Indian accents in spoken Hindi (e.g., Kashmiri, Manipuri, Bengali, neutral

Hindi) [64] or accents within the English language (e.g., American, British, Australian, Scottish, Irish) [45], as well as for the recognition of foreign accents, such as Albanian, Kurdish, Turkish, Arabic and Russian accent in Finnish [9] or Hindi, Russian, Italian, Thai, and Vietnamese accent in English [9, 39].

By means of automated speech analysis, it is not only possible to identify a person's country of origin but also to estimate his or her "degree of nativeness" on a continuous scale [33]. Non-native speakers can even be detected when they are very fluent in the spoken language and have lived in the respective host country for several years [62]. Experimental results show that existing accent recognition systems are effective and have long reached accuracies comparable to human performance [9, 39, 45, 62].

Native language and geographical origin can be sensitive pieces of personal information, which could be misused for the detection and discrimination of minorities. Unfair treatment based on national origin is a widespread form of discrimination [24].

2.9 Speaker Pathology

Through indicative sounds like coughs or sneezes and certain speech parameters, such as loudness, roughness, hoarseness, and nasality, voice recordings may contain rich information about a speaker's state of health [19, 20, 47]. Voice analysis has been described as "one of the most important research topics in biomedical electronics" [104].

Rather obviously, recorded speech may allow inferences about communication disorders, which can be divided into language disorders (e.g., dysphasia, underdevelopment of vocabulary or grammar), voice disorders (e.g., vocal fold paralysis, laryngeal cancer, tracheoesophageal substitute voice) and speech disorders (e.g., stuttering, cluttering) [19, 88].

But also conditions beyond the speech production can be detected from voice samples, including Huntington's disease [76], Parkinson's disease [19], amyotrophic lateral sclerosis [74], asthma [104], Alzheimer's disease [27], and respiratory tract infections caused by the common cold and flu [20]. The sound of a person's voice may even serve as an indicator of overall fitness and long-term health [78, 103].

Further, voice cues may reveal a speaker's smoking habit: A linear relationship has been observed between the number of cigarettes smoked per day and certain voice features, allowing for speech-based smoker detection in a relatively early stage of the habit (<10 years) [30]. Recorded human sounds can also be used for the automatic recognition of physical pain levels [61] and the detection of sleep disorders like obstructive sleep apnea [19].

Computerized methods for speech-based health assessment reach near-human performance in a variety of recognition and analysis tasks and have already been translated into patents [19, 47]. For example, Amazon has patented a system to analyze voice commands recorded by a smart speaker to assess the user's health [47].

The EU's General Data Protection Regulation classifies health-related data as a *special category of personal data* for which particular protection is warranted (Art. 9 GDPR). Among other discriminatory applications, such data may be used by insurance companies to adjust premiums of policyholders according to their state of health [18].

2.10 Mental Health Assessment

Speech abnormalities are a defining characteristic of various mental illnesses. A voice with little pitch variation, for example, is a common symptom in people suffering from schizophrenia or severe depression [36]. Other parameters that may reveal mental health issues include verbal fluency, intonation, loudness, speech tempo, semantic coherence, and speech complexity [8, 31, 36].

Depressive speech can be detected automatically with high accuracy based on voice cues, even under adverse recording conditions, such as low microphone quality, short utterances, and background environmental noise [19, 41]. Not only the detection, but also a severity assessment of depression is possible using a speech sample: In men and women, certain voice features were found to be highly predictive of their HAMD (Hamilton Depression Rating Scale) score, which is the most widely used diagnostic tool to measure a patient's degree of depression and suicide risk [36]. Researchers have even shown that it is possible to predict a future depression based on speech parameters, up to two years before the speaker meets diagnostic criteria [75].

Other mental disorders, such as schizophrenia [31], autism spectrum conditions [19], and post-traumatic stress disorder [102], can also be detected through voice and speech analysis. In some experiments, such methods have already surpassed the classification accuracy of traditional clinical interviews [8].

In common with a person's age, gender, physical health, and national origin, information about mental health problems can be very sensitive, often serving as a basis for discrimination [83].

2.11 Prediction of Interpersonal Perception

A person's voice and manner of expression have a considerable influence on how he or she is perceived by other people [44, 51, 88, 90]. In fact, a single spoken word is enough to obtain personality ratings that are highly consistent across independent listeners [10]. Research has also shown that personality assessments based solely on speech correlate strongly with whole person judgements [88]. Conversely, recorded speech may reveal how a speaker tends to be perceived by other people.

Studies have shown, for example, that fast talkers are perceived as more extroverted, dynamic, and competent [80], that individuals with higher-pitched voices are perceived as more open but less conscientious and emotionally stable [44], that specific intonation patterns increase a speaker's perceived trustworthiness and dominance [81], and that certain prosodic and lexical speech features correlate with observer ratings of charisma [88].

Researchers have also investigated the influence of speech parameters on the perception and treatment of speakers in specific contexts and areas of life. It was found, for instance, that voice cues of elementary school students significantly affect the judgements teachers make about their intelligence and character traits [90]. Similarly, certain speech characteristics of job candidates, including their use of filler words, fluency of speaking, and manner of expression, have been used to predict interviewer ratings for traits such as engagement, excitement, and friendliness [70]. Other studies show that voice plays an important role in the popularity of political candidates as it

influences their perceived competence, strength, physical prowess, and integrity [51]. According to [6], voters tend to prefer candidates with a deeper voice and greater pitch variability. The same phenomenon can be observed in the appointment of board members: CEOs with lower-pitched voices tend to manage larger companies, earn more, and enjoy longer tenures. In [65], a voice pitch decrease of 22.1 Hz was associated with $187 thousand more in annual salary and a $440 million increase in the size of the enterprise managed. On top of this, voice parameters also have a measurable influence on perceived attractiveness and mate choice [44].

Based on voice samples, it is possible to predict how strangers judge a speaker along certain personality traits – a technique referred to as "automatic personality perception" [88]. Considering that the impression people make on others often has a tangible impact on their possibilities and success in life [6, 51, 65, 90], it becomes clear how sensitive and revealing such information can be.

2.12 Inference of Socioeconomic Status

Certain speech characteristics may allow insights into a person's socioeconomic status. There is ample evidence, for instance, that language abilities – including vocabulary, grammatical development, complexity of utterances, productive and receptive syntax – vary significantly between different social classes, starting in early childhood [38]. Therefore, people from distinct socioeconomic backgrounds can often be told apart based on their "entirely different modes of speech" [11]. Besides grammar and vocabulary, researchers found striking inter-class differences in the variety of perspectives utilized in communication and in the use of stylistic devices, observing that once the nature of the difference is grasped, it is "astonishing how quickly a characteristic organization of communication [can] be detected." [87].

Not only language skills, but also the sound of a speaker's voice may be used to draw inferences about his or her social standing. The menarcheal status of girls, for example, which can be derived from voice samples, is used by anthropologists to investigate living conditions and social inequalities in populations [15]. In certain contexts, voice cues, such as pitch and loudness, can even reveal a speaker's hierarchical rank [52].

Based on existing research, it is difficult to say how precise speech-based methods for the assessment of socioeconomic status can become. However, differences between social classes certainly appear discriminative enough to allow for some forms of automatic classification.

2.13 Classification of Acoustic Scenes and Events

Aside from human speech, voice recordings often contain some form of ambient noise. By analyzing background sounds, it is possible to recognize the environment in which an audio sequence was recorded, including indoor environments (e.g., library, restaurant, grocery store, home, metro station, office), outdoor environments (e.g., beach, city center, forest, residential area, urban park), and transport modes (e.g., bus, car, train) [43, 97].

It is also possible to automatically detect and classify specific audio events, such as animal sounds (e.g., dog, cat, crow, crickets), natural sounds (e.g., rain, sea waves, wind, thunderstorm), urban sounds (e.g., church bells, fireworks, jackhammer), office sounds (e.g., mouse click, keyboard typing, printer), bathroom sounds (e.g., showering, urination, defecation, brushing teeth), domestic sounds (e.g., clock tick, page turning, creaking door, keys placed on a table), and non-speech human sounds (e.g., crying, sneezing, breathing, coughing) [4, 16, 43, 97].

Algorithms can even recognize drinking and eating moments in audio recordings and the type of food a person is eating (e.g., soup, rice, apple, nectarine, banana, crisps, biscuits, gummi bears) [19, 91]. Commercial applications like *Shazam* further demonstrate that media sounds, such as songs and movie soundtracks, can be automatically identified and classified into their respective genre with high accuracy, even based on short snippets recorded in a noisy environment [49].

Through such inferences, ambient sounds in audio recordings may not only allow insights into a device holder's context and location, but also into his or her preferences and activities. Certain environments, such as places of worship or street protests, could potentially reveal a person's religious and political affiliations.

Sensitive information can even be extracted from ultrasonic audio signals inaudible to the human ear. An example that has received a lot of media attention recently is the use of so-called "ultrasonic beacons", i.e. high-pitched Morse signals which are secretly emitted by speakers installed in businesses and stores, or embedded in TV commercials and other broadcast content, allowing companies to unobtrusively track the location and media consumption habits of consumers. A growing number of mobile apps – several hundred already, some of them very popular – are using their microphone permission to scan ambient sound for such ultrasonic signals, often without properly informing the user about it [59].

3 Discussion and Implications

As illustrated in the previous section, sensitive inferences can be drawn from human speech and other sounds commonly found in recorded audio. Apart from the linguistic content of a voice recording, a speaker's patterns of word use, manner of pronunciation, and voice characteristics can implicitly contain information about his or her biometric identity, body features, gender, age, personality traits, mental and physical health condition, emotions, intention to deceive, degree of intoxication and sleepiness, geographical origin, and socioeconomic status.

While there is a rich and growing body of research to support the above statement, it has to be acknowledged that many of the studies cited in this paper achieved their classification results under ideal laboratory conditions (e.g., scripted speech, high quality microphones, close-capture recordings, no background noise) [10, 20, 30, 36, 55, 60, 70, 82, 94, 107], which may raise doubt about the generalizability of their inference methods. Also, while impressive accuracies have been reached, it should not be neglected that nearly all of the mentioned approaches still exhibit considerable error rates.

On the other hand, since methods for voice and speech analysis are often subject to non-disclosure agreements, the most advanced know-how arguably rests within the industry and is not publicly available. It can be assumed that numerous corporate and governmental actors with access to speech data from consumer devices possess much larger amounts of training data and more advanced technical capabilities than the researchers cited in this paper. Amazon, for example, spent more than $23 billion on research and development in 2017 alone, has sold more than 100 million Alexa-enabled devices and, according to the company's latest annual report, „customers spoke to Alexa tens of billions more times in 2018 compared to 2017" [108]. Moreover, companies can link speech data with auxiliary datasets (e.g., social media data, browsing behavior, purchase history) to draw other sensitive inferences [47] while the methods considered in this paper exclusively rely on human speech and other sounds commonly found in recorded audio. Looking forward, we expect the risk of unintended information disclosure from speech data to grow further with the continuing proliferation of microphone-equipped devices and the development of more efficient inference algorithms. Deep learning, for instance, still appears to offer significant improvement potential for automated voice analysis [3, 19].

While recognizing the above facts and developments as a substantial privacy threat, it is not our intention to deny the many advantages that speech applications offer in areas like public health, productivity, and convenience. Devices with voice control, for instance, improve the lives of people with physical disabilities and enhance safety in situations where touch-based user interfaces are dangerous to use, e.g., while driving a car. Similarly, the detection of health issues from voice samples (see sect. 2.9) could help in treating illnesses more effectively and reduce healthcare costs.

But since inferred information can be misused in countless ways [17, 18], robust data protection mechanisms are needed in order to reap the benefits of voice and speech analysis in a socially acceptable manner. At the technical level, many approaches have been developed for privacy protection at different stages of the data life cycle, including operations over encrypted data, differential privacy, data anonymization, secure multiparty computation, and privacy-preserving data processing on edge devices [46, 72, 106]. Various privacy safeguards have been specifically designed or adjusted for audio mining applications. These include voice binarization, hashing techniques for speech data, fully homomorphic inference systems, differential private learning, the computation of audio data in separate entrusted units, and speaker de-identification by voice transformation [72, 73]. A comprehensive review of cryptography-based solutions for speech data is provided in [72]. Privacy risks can also be moderated by storing and processing only the audio data required for an application's functionality. For example, where only the linguistic content is required, voice recordings can be converted to text in order to eliminate all voice-related information and thereby minimize the potential for undesired inferences.

In advocating data collection transparency and informational self-determination, the recent privacy discourse has put a focus on the recording mode of microphone-equipped devices, where a distinction can be made between "manually activated," "speech activated," and "always on" [34]. However, data scandals show that reporting modes cannot always be trusted [105]. And even where audio is only recorded and transmitted

with a user's explicit consent, sensitive inferences may unnoticeably be drawn from collected speech data, ultimately leaving the user without control over his or her privacy. Enabling the unrestricted screening of audio data for potentially revealing patterns and correlations, recordings are often available to providers of cloud-based services in unencrypted form – an example being voice-based virtual assistants [1, 22]. With personal data being the foundation for highly profitable business models and strategic surveillance practices, it is certainly not unusual for speech data to be processed in an unauthorized or unexpected manner. This is well illustrated by recently exposed cases where Amazon, Google, and Apple ordered human contractors to listen to private voice recordings of their customers [22].

The findings compiled in this paper reveal a serious threat to consumer privacy and show that more research is needed into the societal implications of voice and speech processing. In addition to investigating the technical feasibility of inferences from speech data in more detail, future research should explore technical and legal countermeasures to the presented problem, including ways to enforce existing data protection laws more effectively. Of course, the problem of undesired inferences goes far beyond microphones and needs to be addressed for other data sources as well. For example, in recent work, we have also investigated the wealth of sensitive information that can be implicitly contained in data from air quality sensors, infrared motion detectors, smart meters [56], accelerometers [57], and eye tracking sensors [58]. It becomes apparent that sensors in many everyday electronic devices can reveal significantly more information than one would assume based on their advertised functionality. The crafting of solutions to either limit the immense amounts of knowledge and power this creates for certain organizations, or to at least avert impending negative consequences, will be an important challenge for privacy, social justice, and civil rights advocates over the years to come.

4 Conclusion

Microphones are widely used in connected devices, where they have a large variety of possible applications. While recognizing the benefits of voice and speech analysis, this paper highlights the growing privacy threat of unexpected inferences from audio data. Besides the linguistic content, a voice recording can implicitly contain information about a speaker's identity, personality, body shape, mental and physical health, age, gender, emotions, geographical origin, and socioeconomic status – and may thereby potentially reveal much more information than a speaker wishes and expects to communicate.

Further research is required into the privacy implications of microphone-equipped devices, taking into account the evolving state of the art in data mining technology. As it is impossible, however, to meaningfully determine the limits of inference methods developed behind closed doors, voice recordings – even where the linguistic content does not seem rich and revealing – should be regarded and treated as highly sensitive by default. Since existing technical and legal countermeasures are limited and do not yet offer reliable protection against large-scale misuses of audio data and undesired

inferences, more effective safeguards and means of enforcement are urgently needed. We hope that the knowledge compiled in this paper can serve as a basis for consumer education and will help lawmakers and fellow researchers in assessing the richness and potential sensitivity of speech data.

References

1. Amazon: Alexa and Alexa Device FAQs, https://www.amazon.com/gp/help/customer/display.html?nodeId=201602230, last accessed 2019/11/16.
2. An, G. et al.: Deep Personality Recognition for Deception Detection. In: INTERSPEECH. pp. 421–425 (2018). https://doi.org/10.21437/Interspeech.2018-2269.
3. An, G., Levitan, R.: Lexical and Acoustic Deep Learning Model for Personality Recognition. In: INTERSPEECH. pp. 1761–1765 (2018).
4. Aytar, Y. et al.: SoundNet: Learning Sound Representations from Unlabeled Video. In: Conference on Neural Information Processing Systems (NIPS). pp. 892–900 (2016).
5. Bae, S.-G. et al.: A Judgment of Intoxication using Hybrid Analysis with Pitch Contour Compare in Speech Signal Processing. IJAER. 12, 10, 2342–2346 (2017).
6. Banai, B. et al.: Candidates' voice in political debates and the outcome of presidential elections. In: Psychology Days in Zadar. pp. 33–39 University of Zadar (2017).
7. Bedi, G. et al.: A Window into the Intoxicated Mind? Speech as an Index of Psychoactive Drug Effects. Neuropsychopharmacology. 39, 10, 2340–2348 (2014).
8. Bedi, G. et al.: Automated analysis of free speech predicts psychosis onset in high-risk youths. Npj Schizophr. 1, 15030 (2015).
9. Behravan, H. et al.: i-Vector Modeling of Speech Attributes for Automatic Foreign Accent Recognition. Trans. Audio Speech Lang. Process. 24, 1, 29–41 (2016).
10. Belin, P. et al.: The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. PLOS ONE. 12, 10, e0185651 (2017).
11. Bernstein, B.: Language and Social Class. Br. J. Sociol. 11, 3, 271–276 (1960).
12. Bindahman, S. et al.: 3D body scanning technology: Privacy and ethical issues. In: Conference on Cyber Security, Cyber Warfare and Digital Forensic. pp. 150–154 (2012).
13. Bone, D. et al.: Intoxicated Speech Detection. Comput. Speech Lang. 28, 2, 375–391 (2014). https://doi.org/10.1016/j.csl.2012.09.004.
14. Booth, R.: Facebook reveals news feed experiment to control emotions, https://www.theguardian.com/technology/2014/jun/29/facebook-users-emotions-news-feeds, (2014).
15. Bugdol, M.D. et al.: Prediction of menarcheal status of girls using voice features. Comput. Biol. Med. 100, 296–304 (2018). https://doi.org/10.1016/j.compbiomed.2017.11.005.
16. Chen, J. et al.: Bathroom Activity Monitoring Based on Sound. In: Gellersen, H.-W. et al. (eds.) Pervasive Computing. pp. 47–61 Springer, Berlin (2005).
17. Christl, W.: How Companies Use Data Against People. Cracked Labs, Vienna (2017).
18. Christl, W., Spiekermann, S.: Networks of Control: A Report on Corporate Surveillance, Digital Tracking, Big Data & Privacy. Facultas, Vienna (2016).
19. Cummins, N. et al.: Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. Methods. 151, 41–54 (2018).
20. Cummins, N. et al.: "You sound ill, take the day off": Automatic recognition of speech affected by upper respiratory tract infection. In: IEEE EMBC. pp. 3806–3809 (2017).
21. Desplanques, B., Demuynck, K.: Cross-lingual Speech Emotion Recognition through Factor Analysis. In: INTERSPEECH. pp. 3648–3652 (2018).

22. Drozdiak, N., Turner, G.: Apple, Google, and Amazon May Have Violated Your Privacy by Reviewing Digital Assistant Commands, https://fortune.com/2019/08/05/google-apple-amazon-digital-assistants/, last accessed 2019/09/03.
23. Dubey, H. et al.: BigEAR: Inferring the Ambient and Emotional Correlates from Smartphone-based Acoustic Big Data. In: IEEE CHASE. pp. 78–83 (2016).
24. EEOC: Charge Statistics, https://www.eeoc.gov/eeoc/statistics/enforcement/charges.cfm, last accessed 2019/11/07.
25. Evans, S. et al.: Relationships between vocal characteristics and body size and shape in human males. Biol. Psychol. 72, 2, 160–163 (2006).
26. Fast, L.A., Funder, D.C.: Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. J. Pers. Soc. Psychol. 94, 2, 334–346 (2008).
27. Fraser, K.C. et al.: Linguistic Features Identify Alzheimer's Disease in Narrative Speech. J. Alzheimers Dis. 49, 2, 407–422 (2015). https://doi.org/10.3233/JAD-150520.
28. Ghahremani, P. et al.: End-to-end Deep Neural Network Age Estimation. In: INTERSPEECH. pp. 277–281 (2018). https://doi.org/10.21437/Interspeech.2018-2015.
29. González, J.: Correlations between speakers' body size and acoustic parameters of voice. Percept. Mot. Skills. 105, 1, 215–220 (2007).
30. Gonzalez, J., Carpi, A.: Early effects of smoking on the voice: a multidimensional study. Med. Sci. Monit. 10, 12, CR649-56 (2004).
31. Gosztolya, G. et al.: Identifying Schizophrenia Based on Temporal Parameters in Spontaneous Speech. In: INTERSPEECH. pp. 3408–3412 (2018).
32. Gray, S.: Always On: Privacy Implications of Microphone-Enabled Devices. Future of Privacy Forum, Washington, DC (2016).
33. Grosz, T. et al.: Assessing the Degree of Nativeness and Parkinson's Condition Using Gaussian Processes and Deep Rectifier Neural Networks. In: INTERSPEECH. (2015).
34. Grzybowska, J., Ziółko, M.: I-vectors in gender recognition from telephone speech. In: Conference on Applications of Mathematics in Biology and Medicine. (2015).
35. Haider, F. et al.: An Active Feature Transformation Method for Attitude Recognition of Video Bloggers. In: INTERSPEECH. pp. 431–435 (2018).
36. Hashim, N.W. et al.: Evaluation of Voice Acoustics as Predictors of Clinical Depression Scores. J. Voice. 31, 2, 256.e1-256.e6 (2017). https://doi.org/10.1016/j.jvoice.2016.06.006.
37. Herms, R.: Prediction of Deception and Sincerity from Speech Using Automatic Phone Recognition-Based Features. In: INTERSPEECH. pp. 2036–2040 (2016).
38. Hoff, E.: How social contexts support and shape language development. Dev. Rev. 26, 1, 55–88 (2006). https://doi.org/10.1016/j.dr.2005.11.002.
39. Honig, F. et al.: Islands of Failure: Employing word accent information for pronunciation quality assessment of English L2 learners. In: ISCA SLATE Workshop. (2009).
40. HSBC: Welcome to Voice ID, https://www.us.hsbc.com/customer-service/voice/, last accessed 2019/10/22.
41. Huang, Z. et al.: Depression Detection from Short Utterances via Diverse Smartphones in Natural Environmental Conditions. In: INTERSPEECH. pp. 3393–3397 (2018).
42. Hughes, S.M. et al.: Sex-specific body configurations can be estimated from voice samples. J. Soc. Evol. Cult. Psychol. 3, 4, 343–355 (2009). https://doi.org/10.1037/h0099311.
43. IEEE AASP: Challenge results published, http://www.cs.tut.fi/sgn/arg/dcase2017/articles/challenge-results-published, last accessed 2019/10/22.
44. Imhof, M.: Listening to Voices and Judging People. Int. J. List. 24, 1, 19–33 (2010).
45. Jain, A. et al.: Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning. In: INTERSPEECH. pp. 2454–2458 (2018).

46. Jain, P. et al.: Big data privacy: a technological perspective and review. J. Big Data. 3, 1, 25 (2016). https://doi.org/10.1186/s40537-016-0059-y.
47. Jin, H., Wang, S.: Voice-based determination of physical and emotional characteristics of users, https://patents.google.com/patent/US10096319B1/en?oq=10096319, (2018).
48. Kabil, S.H. et al.: On Learning to Identify Genders from Raw Speech Signal Using CNNs. In: INTERSPEECH. pp. 287–291 (2018).
49. Kaneshiro, B. et al.: Characterizing Listener Engagement with Popular Songs Using Large-Scale Music Discovery Data. Front. Psychol. 8, 1–15 (2017).
50. Karpey, D., Pender, M.: Customer Identification Through Voice Biometrics, https://patents.google.com/patent/US9396730, (2016).
51. Klofstad, C.A. et al.: Perceptions of Competence, Strength, and Age Influence Voters to Select Leaders with Lower-Pitched Voices. PLOS ONE. 10, 8, e0133779 (2015).
52. Ko, S.J. et al.: The Sound of Power: Conveying and Detecting Hierarchical Rank Through Voice. Psychol. Sci. 26, 1, 3–14 (2015). https://doi.org/10.1177/0956797614553009.
53. Koolagudi, S.G. et al.: IITKGP-SESC: Speech Database for Emotion Analysis. In: Ranka, S. et al. (eds.) Contemporary Computing. pp. 485–492 Springer, Berlin (2009).
54. Kotenko, J.: To infinity and Beyond Verbal, https://www.digitaltrends.com/social-media/exploring-beyond-verbal-the-technology-of-emotions-analytics/, (2013).
55. Krauss, R.M. et al.: Inferring speakers' physical attributes from their voices. J. Exp. Soc. Psychol. 38, 6, 618–625 (2002).
56. Kröger, J.: Unexpected Inferences from Sensor Data: A Hidden Privacy Threat in the Internet of Things. In: Strous, L. and Cerf, V.G. (eds.) Internet of Things. Information Processing in an Increasingly Connected World. pp. 147–159 Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-15651-0_13.
57. Kröger, J.L. et al.: Privacy Implications of Accelerometer Data: A Review of Possible Inferences. In: Proceedings of the 3rd International Conference on Cryptography, Security and Privacy (ICCSP). ACM, New York (2019). https://doi.org/10.1145/3309074.3309076.
58. Kröger, J.L. et al.: What Does Your Gaze Reveal About You? On the Privacy Implications of Eye Tracking. In: Fricker, S. et al. (eds.) Privacy and Identity Management. Springer, Cham (2019).
59. Kröger, J.L., Raschke, P.: Is My Phone Listening in? On the Feasibility and Detectability of Mobile Eavesdropping. In: Foley, S.N. (ed.) Data and Applications Security and Privacy XXXIII. pp. 102–120 Springer (2019). https://doi.org/10.1007/978-3-030-22479-0_6.
60. Levitan, S.I. et al.: Acoustic-Prosodic Indicators of Deception and Trust in Interview Dialogues. In: INTERSPEECH. pp. 416–420 (2018).
61. Li, J.-L. et al.: Learning Conditional Acoustic Latent Representation with Gender and Age Attributes for Automatic Pain Level Recognition. In: INTERSPEECH. (2018).
62. Lopes, J. et al.: A nativeness classifier for TED Talks. In: ICASSP. pp. 5672–5675 (2011).
63. Magi, T.J.: Fourteen Reasons Privacy Matters: A Multidisciplinary Review of Scholarly Literature. Libr. Q. Inf. Community Policy. 81, 2, 187–209 (2011).
64. Malhotra, K., Khosla, A.: Automatic identification of gender & accent in spoken Hindi utterances with regional Indian accents. In: IEEE SLT Workshop. pp. 309–312 (2008).
65. Mayew, W.J. et al.: Voice pitch and the labor market success of male chief executive officers. Evol. Hum. Behav. 34, 4, 243–248 (2013).
66. McLaren, M. et al.: The 2016 Speakers in the Wild Speaker Recognition Evaluation. In: INTERSPEECH. pp. 823–827 (2016). https://doi.org/10.21437/Interspeech.2016-1137.
67. Mendels, G. et al.: Hybrid Acoustic-Lexical Deep Learning Approach for Deception Detection. In: INTERSPEECH. pp. 1472–1476 (2017).

68. Mohammadi, G. et al.: The voice of personality: mapping nonverbal vocal behavior into trait attributions. In: Workshop on Social Signal Processing (SSPW). pp. 17–20 (2010).
69. Mporas, I., Ganchev, T.: Estimation of unknown speaker's height from speech. Int. J. Speech Technol. 12, 4, 149–160 (2009). https://doi.org/10.1007/s10772-010-9064-2.
70. Naim, I. et al.: Automated prediction and analysis of job interview performance. In: IEEE Conference on Automatic Face and Gesture Recognition. pp. 1–6 (2015).
71. Nandwana, M.K. et al.: Robust Speaker Recognition from Distant Speech under Real Reverberant Environments Using Speaker Embeddings. In: INTERSPEECH. (2018).
72. Nautsch, A. et al.: Preserving privacy in speaker and speech characterisation. Comput. Speech Lang. 58, 441–480 (2019). https://doi.org/10.1016/j.csl.2019.06.001.
73. Nautsch, A. et al.: The GDPR & Speech Data: Reflections of Legal and Technology Communities, First Steps towards a Common Understanding. In: INTERSPEECH. pp. 3695–3699 (2019). https://doi.org/10.21437/Interspeech.2019-2647.
74. Norel, R. et al.: Detection of Amyotrophic Lateral Sclerosis (ALS) via Acoustic Analysis. In: INTERSPEECH. pp. 377–381 (2018). https://doi.org/10.1101/383414.
75. Ooi, K.E.B. et al.: Multichannel Weighted Speech Classification System for Prediction of Major Depression in Adolescents. IEEE Trans. Biomed. Eng. 60, 2, 497–506 (2013).
76. Perez, M. et al.: Classification of Huntington Disease Using Acoustic and Lexical Features. In: INTERSPEECH. pp. 1898–1902 (2018).
77. Petrushin, V.A.: Detecting emotions using voice signal analysis, https://patents.google.com/patent/US7222075B2/en, (2007).
78. Pipitone, R.N., Gallup, G.G.: Women's voice attractiveness varies across the menstrual cycle. Evol. Hum. Behav. 29, 4, 268–274 (2008).
79. Polzehl, T. et al.: Automatically Assessing Personality from Speech. In: IEEE Conference on Semantic Computing (ICSC). pp. 134–140 (2010).
80. Polzehl, T.: Personality in Speech. Springer International Publishing, Cham (2015).
81. Ponsot, E. et al.: Cracking the social code of speech prosody using reverse correlation. Proc. Natl. Acad. Sci. 115, 15, 3972–3977 (2018).
82. Ranganath, R. et al.: Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. Comput. Speech Lang. 27, 1, 89–115 (2013).
83. Reavley, N.J., Jorm, A.F.: Experiences of discrimination and positive treatment in people with mental health problems. Aust. N. Z. J. Psychiatry. 49, 10, 906–913 (2015).
84. Reubold, U. et al.: Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers. Speech Commun. 52, 7–8, 638–651 (2010).
85. Sadjadi, S.O. et al.: Speaker age estimation on conversational telephone speech using senone posterior based i-vectors. In: ICASSP. pp. 5040–5044 (2016).
86. Sarma, M. et al.: Emotion Identification from Raw Speech Signals Using DNNs. In: INTERSPEECH. pp. 3097–3101 (2018). https://doi.org/10.21437/Interspeech.2018-1353.
87. Schatzman, L., Strauss, A.: Social Class and Modes of Communication. Am. J. Sociol. 60, 4, 329–338 (1955). https://doi.org/10.1086/221564.
88. Schuller, B. et al.: A Survey on perceived speaker traits: Personality, likability, pathology, and the first challenge. Comput. Speech Lang. 29, 1, 100–131 (2015).
89. Schuller, B. et al.: Medium-term speaker states - A review on intoxication, sleepiness and the first challenge. Comput. Speech Lang. 28, 2, 346–374 (2013).
90. Seligman, C.R. et al.: The effects of speech style and other attributes on teachers' attitudes toward pupils. Lang. Soc. 1, 01, 131 (1972).
91. Sim, J.M. et al.: Acoustic Sensor Based Recognition of Human Activity in Everyday Life for Smart Home Services. Int. J. Distrib. Sens. Netw. 11, 9, 679123 (2015).

92. Simpson, A.P.: Phonetic differences between male and female speech. Lang. Linguist. Compass. 3, 2, 621–640 (2009). https://doi.org/10.1111/j.1749-818X.2009.00125.x.
93. Sloboda, J. et al.: Vocal Biomarkers for Cognitive Performance Estimation in a Working Memory Task. In: INTERSPEECH. pp. 1756–1760 (2018).
94. Soskin, W.F., Kauffman, P.E.: Judgment of Emotion in Word-Free Voice Samples. J. Commun. 11, 2, 73–80 (1961). https://doi.org/10.1111/j.1460-2466.1961.tb00331.x.
95. Stanek, M., Sigmund, M.: Psychological Stress Detection in Speech Using Return-to-opening Phase Ratios in Glottis. Elektron Elektrotech. 21, 5, 59–63 (2015).
96. Stanescu, C.G., Ievchuk, N.: Alexa, Where Is My Private Data? In: Digitalization in Law. pp. 237–247 Social Science Research Network, Rochester, NY (2018).
97. Stowell, D. et al.: Detection and Classification of Acoustic Scenes and Events. IEEE Trans. Multimed. 17, 10, 1733–1746 (2015). https://doi.org/10.1109/TMM.2015.2428998.
98. Streeter, L.A. et al.: Pitch changes during attempted deception. J. Pers. Soc. Psychol. 35, 5, 345–350 (1977). https://doi.org/10.1037//0022-3514.35.5.345.
99. Swain, M. et al.: Databases, features and classifiers for speech emotion recognition: a review. Int. J. Speech Technol. 21, 1, 93–120 (2018).
100. Trilok, N.P. et al.: Establishing the Uniqueness of the Human Voice for Security Applications. In: Proceedings of Student-Faculty Research Day. pp. 8.1-8.6 Pace University (2004).
101. Tsai, F.-S. et al.: Automatic Assessment of Individual Culture Attribute of Power Distance Using a Social Context-Enhanced Prosodic Network Representation. In: INTERSPEECH. pp. 436–440 (2018). https://doi.org/10.21437/Interspeech.2018-1523.
102. Vergyri, D. et al.: Speech-Based Assessment of PTSD in a Military Population Using Diverse Feature Classes. In: INTERSPEECH. pp. 3729–3733 (2015).
103. Vukovic, J. et al.: Women's voice pitch is negatively correlated with health risk factors. J. Evol. Psychol. 8, 3, 217–225 (2010). https://doi.org/10.1556/JEP.8.2010.3.2.
104. Walia, G.S., Sharma, R.K.: Level of asthma: Mathematical formulation based on acoustic parameters. In: CASP. pp. 24–27 (2016). https://doi.org/10.1109/CASP.2016.7746131.
105. Wolfson, S.: Amazon's Alexa recorded private conversation and sent it to random contact, https://www.theguardian.com/technology/2018/may/24/amazon-alexa-recorded-conversation, (2018).
106. Zhao, J. et al.: Privacy-Preserving Machine Learning Based Data Analytics on Edge Devices. In: AAAI/ACM Conference on AI, Ethics, and Society (AIES). pp. 341–346 (2018).
107. Zhou, Y. et al.: Deception detecting from speech signal using relevance vector machine and non-linear dynamics features. Neurocomputing. 151, 1042–1052 (2015).
108. Annual Report 2018. Amazon.com, Inc., Seattle, Washington, USA (2019).