IEICE TRANS. INF. & SYST., VOL.E104–D, NO.1 JANUARY 2021

INVITED PAPER *Special Section on Enriched Multimedia — Multimedia Security and Forensics —*

# Preventing Fake Information Generation Against Media Clone Attacks**

**Noboru BABAGUCHI**[†a)], **Isao ECHIZEN**[††], ***Fellows***, **Junichi YAMAGISHI**[††], **Naoko NITTA**[†], **Yuta NAKASHIMA**[†], **Kazuaki NAKAMURA**[†], **Kazuhiro KONO**[†††], ***Members***, **Fuming FANG**[††], ***Nonmember***, **Seiko MYOJIN**[†], ***Member***, **Zhenzhong KUANG**[†*], **Huy H. NGUYEN**[††], ***and*** **Ngoc-Dung T. TIEU**[††], ***Nonmembers***

**SUMMARY** Fake media has been spreading due to remarkable advances in media processing and machine leaning technologies, causing serious problems in society. We are conducting a research project called Media Clone aimed at developing methods for protecting people from fake but skillfully fabricated replicas of real media called media clones. Such media can be created from fake information about a specific person. Our goal is to develop a trusted communication system that can defend against attacks of media clones. This paper describes some research results of the Media Clone project, in particular, various methods for protecting personal information against generating fake information. We focus on 1) fake information generation in the physical world, 2) anonymization and abstraction in the cyber world, and 3) modeling of media clone attacks.

***key words:*** *media clones, fake media, media processing, anonymization, trusted communication*

## 1. Introduction

The remarkable advances in media processing and machine learning technologies in recent years have reduced the distance between real and fake media. We refer to media that expresses the real world as it as *real media* and media that is enhanced using media technologies, such as computer graphics and voice conversion, as *fake media*. Distribution of realistic fake media has become a threat to our daily lives as exemplified by the use of voice spoofing to commit telephone-based fraud against family members or friends. The more the voice in the fake media resembles that of the actual person, the greater the risk of the fraud being successful.

To achieve a safe and reliable society, it is of great importance to protect people against fake but skillfully fabricated replicas of real media, called *media clones*, by means of media processing technologies. In 2016, we launched the Media Clone (MC) research project, aimed at developing a trusted communication system that can defend against attacks of media clones [1].

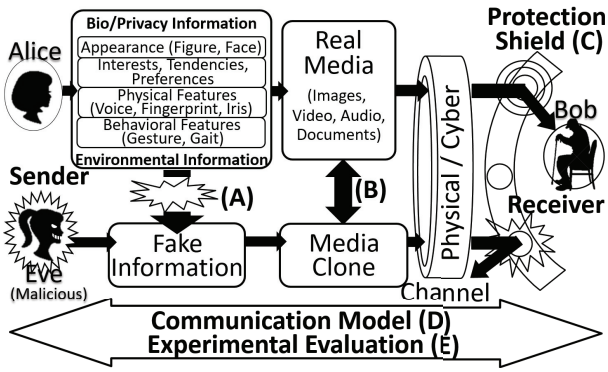



**Fig. 1** Framework of MC project.

Figure 1 shows the framework of the MC project. A sender Alice sends real media such as video or audio to a receiver Bob through physical and cyber channels. At that time, a malicious attacker Eve stealthily acquires privacy, biological, and/or environmental information about Alice to use in creating fake information. Using that fake information, Eve generates clones of Alice's media and sends these media clones to Bob to deceive him.

To create a trusted communication system for defending against attacks by media clones, we are pursuing five themes in the MC project as illustrated in Fig. 1. (A) Development of methods for protecting privacy, biological, and environmental information to prevent fake information generation. (B) Verification of the ability to generate various types of media clones for audio, visual, and text derived from fake information. (C) Realization of a shield for protecting against media clone attacks by recognizing them. (D) Modeling of a trusted communication system that can defend against media clone attacks. (E) Experimental evaluation including regular spoofing versus anti-spoofing competitions. This paper presents the research results of themes (A) and (D). The results of themes (B), (C) and (E) are described in [2].

## 2. Related Work

Abusive use of media clones in both the physical and cyber worlds may lead to serious consequences, including identity theft, and its prevention has become a major concern in

Manuscript received April 7, 2020.
Manuscript revised July 2, 2020.
Manuscript publicized October 19, 2020.

[†]The authors are with Osaka University, Suita-shi, 565–0871 Japan.
[††]The authors are with National Institute of Informatics, Tokyo, 101–8430 Japan.
[†††]The author is with Kansai University, Takatsuki-shi, 569–1098 Japan.
[*]Presently, with Hangzhou Dianzi University, Hangzhou, Zhejiang, China.

[**]This work was supported in part by JSPS KAKENHI Grant Numbers JP16H06302, JP18H04120, and JST CREST Grant Number JPMJCR18A6, Japan.

a) E-mail: babaguchi@comm.eng.osaka-u.ac.jp

DOI: 10.1587/transinf.2020MUI0001

society.

### 2.1 Physical World

One way to prevent media cloning in the *physical world* is to disturb identity cues and/or biometrics using special devices or body attachments. Several methods for preventing recognition/detection of a person's face, speech, and/or appearance were explored even prior to the MC project.

Three methods have been proposed for preventing identification of a person based on facial recognition. One uses patterned coloring of the hair and special paint patterns on the face to cause face detection to fail [3]. Another method uses a device worn on the face to transmit near-infrared signals that are picked up even by RGB cameras, which makes the face undetectable [4], [5]. The third method uses printed eyeglass frames that enable the wearer to evade recognition or impersonate another individual [6].

For speech, a simple method is to add background noise to the speech waveform in the physical space so that speaker-related sensitive information cannot be identified from the speech. A more sophisticated method is to generate noises that degrade speaker verification performance without degrading speech intelligibility [7].

Several adversarial-example-based methods have been proposed to protect human appearance. One uses rigid printable adversarial patches as cloaking devices to fool human detectors [8]. Another method uses non-rigid printable adversarial patches to prevent deep neural network (DNN)-based person detectors from detecting moving people in a video [9]. This method illustrates the potential of applying adversarial perturbations to human clothing, which exhibits non-rigid deformation.

### 2.2 Cyber World

Various signals from a person (face, speech, gait, personal preferences, etc.) can be easily captured using cameras, microphones, and other sensors and digitized. Due to the inherent nature of digitized data, generating media clones is much easier in the cyber world than in the physical one. This has resulted in an enormous amount of research on preventing media cloning for various modalities, especially face, speech, and gait.

Several techniques have been proposed for anonymizing faces. Classic approaches use such image processing techniques as blocking-out, blurring, and pixelation [10]–[15]. The results may not necessarily be natural or sufficiently anonymized [16]–[19]. The concept of $k$-anonymity [20] provides a theoretical guarantee of privacy [21]–[25]. Neural networks or, more specifically, generative adversarial networks (GANs), offer a new approach to anonymizing faces by synthesizing realistic yet anonymous faces [17], [26]–[29], which can be used together with $k$-anonymity or other theoretical criteria [25], [30].

For speech, several classic approaches for speaker anonymization are explored in [31]–[34], while a neural-network-based encoder-decoder approach has been proposed in [35]. A reversible variant can be also found [36]. Methods based on these approaches mainly transform the original speaker's identity into someone else's identity. Other factors, such as speech quality, naturalness, and linguistic information, should also be considered so that the anonymized speech can be used for further analysis.

The use of a person's gait taken from a video sequence has recently been studied for use in identification [37]–[39]. Despite its potential, few studies have been reported on gait anonymization. Although image processing techniques used for face anonymization can also be used for gait anonymization, a more sophisticated approach has been proposed [40]. 3D feature analysis has been done to unveil essential signals for identification, which in turn is beneficial for anonymization [41].

## 3. Fake Information Generation in Physical World

In this section, we take up the threat of fake information generation occurring in the physical world and outline our achievements. Specifically, we discuss the theft of fingerprint information by photography and measures to prevent it as well as playback attacks based on speech enhancement.

### 3.1 Theft of Fingerprint Information by Photography and Countermeasures [42], [43]

Biometric authentication has become widespread and is often implemented as a standard feature of personal devices such as personal computers and smartphones. The spread of fingerprint authentication is particularly remarkable. Along with this trend has come image sensors with higher resolution. This has resulted in concern about fingerprint information that could only be read with a conventional contact-based fingerprint sensor being remotely captured and stolen. In 2014, a German hacker announced that he had successfully recovered fingerprint information from photographs of the German Defense Minister's fingers taken with a commercially available digital camera [44]. Since biometric information used for biometric authentication is immutable for life, once leaked, it poses a threat of spoofing and may pose a significant disadvantage over the lifetime of the victim.

It has been shown that fingerprint information can be extracted from photos at a level sufficient for authentication [42]. In addition, a method has been proposed that prevents the extraction of fingerprint information from a fingerprint image [42]. Specifically, a jamming pattern composed of a pseudo-fingerprint and attached to a persons fingertips prevents extraction of fingerprint information from a photographed image while still enabling the use of contact-based fingerprint sensors.

Figure 2 shows the interference effect of the proposed method during photography. The upper images are for the proposed method, and the lower ones are for the previous method [43]. In the previous method, the feature points

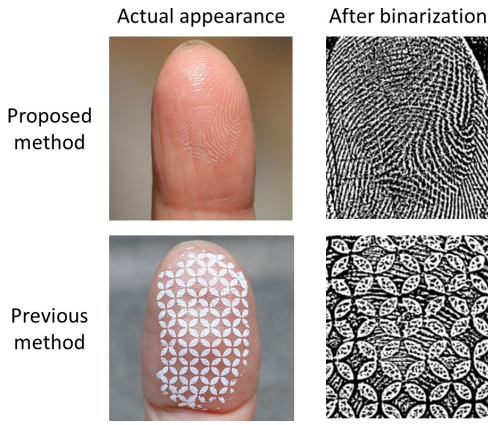



**Fig. 2** Effect of jamming pattern on photographed fingerprint.

of the fingerprint are concealed by emphasizing the edges of a geometric pattern, whereas in the proposed method, a fake feature point is created by superimposing a pseudo-fingerprint pattern on the original fingerprint. In the previous method, the superimposed geometric pattern is periodic, and it is expected that an attack method removes the pattern by frequency separation or weakens the interference effect by filling the pattern part with the skin color. In the proposed method, the superimposed pattern is a pseudo-fingerprint made of a translucent material, making it difficult to discriminate between the actual fingerprint and the pseudo-fingerprint. There is almost no noise superposition due to the jamming pattern when the fingerprint is read by a contact-based fingerprint sensor, so matching of the fingerprint with the registered fingerprint is not affected.

### 3.2 Playback Attacks Based on Speech Pre-Enhancement [45]

Automatic speaker verification (ASV) identifies a person from a set of speech recordings. It is widely used in electronic devices such as mobile phones and smart speakers. However, ASV systems are vulnerable to speech that is recorded and played back at the time of authentication. This is because played-back speech contains the same person-specific characteristics as the genuine speech, A countermeasure (CM) is thus used to filter out replayed speech in order to avoid false acceptance.

In this subsection, we introduce a new attack method that is effective against replayed speech countermeasures. Since the replayed speech contains more noise and reverberation than natural/genuine speech, we can transform the speech into a form that is close to natural speech by using a speech transformation method (e.g., speech enhancement) before replaying the speech [45] to spoof the countermeasure. We carried out an experiment using genuine speech used in the ASVspoof 2017 challenge and used a speech enhancement generative adversarial network (SEGAN) [46] to reduce noise and reverberation. We evaluated this method using two CMs tested at the challenge: the baseline method [47] (a Gaussian-mixture model (GMM)-universal background model (UBM) method) and the best method (a Laplacian convolutional neural network (L-CNN) deep-learning-based method).

To evaluate how many genuine speech samples and how many replayed speech samples were not detected by the CMs, we split the genuine speech data equally into two subsets: one to be used as natural speech and the other to be used as *spoofed speech*. We replayed the spoofed speech and recorded it again as *non-enhanced replay speech*. We then enhanced the spoofed speech, replayed it, and recorded it again as *enhanced played-back speech*. For the baseline CM, the equal error rate (EER) was 4%-10% higher with the enhanced replayed speech compared with the non-enhanced replayed speech. For the L-CNN CM, the EER was 1%-2% higher. We also tested the CMs with a GMM-UBM-based ASV system using a tandem detection cost function (t-DCF) [48]. The t-DCF was higher with the enhanced replayed speech. This indicates that even stronger countermeasures are required for ASV systems.

## 4. Anonymization and Abstraction in Cyber World

Nowadays many people upload and share their media data in the cyber world, especially through social networks. Such data are at risk of being maliciously exploited to generate fake information. Hence, methods for anonymizing and abstracting various types of media data are needed. Since people share media data through social networks in order to view it together with their friends, families, and so on, anonymization methods yielding unnatural results (e.g., blocking-out faces in a video) are undesirable as they would likely annoy viewers. Therefore, it is important not only to remove identities from media data but also to keep their naturalness. We are working on methods for anonymizing different types of media data without losing their naturalness.

### 4.1 Face Anonymization with $k$-Anonymity

Face anonymization has been one of the main social concerns in the social networking era. The situation has been greatly complicated by the development and enhancement of deepfake-related technologies [49], [50], some of which can synthesize a talking head even from a single face image. Face anonymization is thus becoming a fundamental technology for preventing identity theft as well as protecting privacy.

Our approach to face anonymization utilizes $k$-anonymity [20] to generate an anonymized face for an input image from $k$ identities (the *$k$-anonymity identity group*) retrieved from a pool of faces using a GAN framework as shown in Fig. 3. To maintain the utility of the input face to be anonymized so that some facial attributes (age, gender, etc.) are preserved, we use $k$ identities proximate to the input face in the face embedding space obtained using encoder $E$, which is trained on the face pool. An anonymized face is generated using generator $G$. For train-

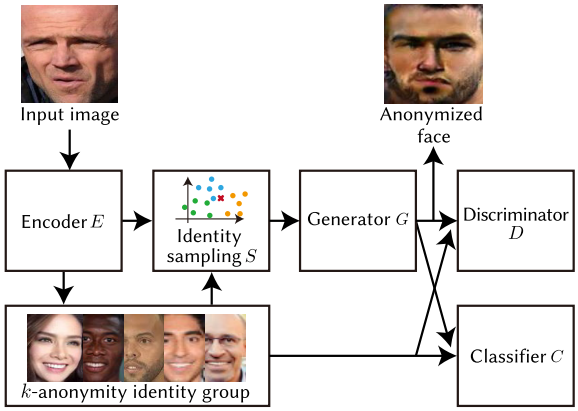



**Fig. 3** Network architecture for face anonymization with $k$-anonymity.

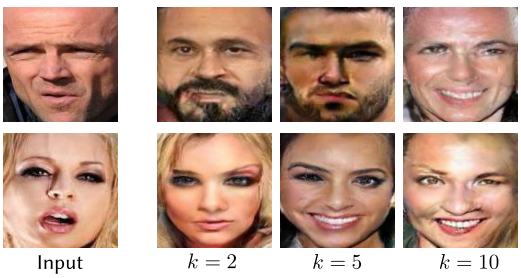



**Fig. 4** Examples of anonymized faces for different $k$'s.

ing, generated faces are fed to discriminator $D$ for computing the adversarial loss and to classifier $C$ for ensuring that the anonymized face is not classified as any face in the face pool. In addition to using the adversarial and classifier-based losses, we use two different losses in the face embedding space so that the anonymized face satisfies $k$-anonymity.

Figure 4 shows examples of anonymized faces for the input image at the left using three values of $k$. Although a larger $k$ tends to result in a collapsed face, some of the facial attributes were preserved. A quantitative evaluation showed that the anonymized faces were not recognized as either the input face's identity or any of the $k$ identities, which is supported by our subjective evaluation.

### 4.2 Neural Speaker Anonymization [51]

Speech data contain much sensitive information that can be used to identify the speaker. In this subsection, we introduce our initial work on speaker anonymization [51]. The speaker anonymization technique we have developed can conceal a speaker's identity while keeping other factors such as linguistic information, naturalness, and quality unchanged. Figure 5 shows our initial system. Pitch, linguistic information, and speaker identity are extracted from the user's speech. The fundamental frequency (F0) is used as an indicator of pitch, and a DNN-based automatic speech recognition (ASR) system [52] is used to extract the linguistic feature representations, i.e., a phoneme posteriorgram (PPG). An x-vector [53], which is widely used in modern ASV systems, is used to represent speaker identity. Speaker anonymization is achieved by replacing the original speaker's x-vector with an anonymous x-vector composed by averaging a set of x-vectors selected from an x-vector pool for a large number of external speakers (7325 speakers). This method enables any type of pseudo-speaker to be easily generated by changing the distance between the original x-vector and the composed x-vector. Next, we learn an acoustic model that transforms the original F0 and PPG and the anonymized x-vector into the pseudo-speaker's acoustic feature mel-spectrogram. Finally, the pseudo-speaker's speech is synthesized from the mel-spectrogram using a neural source filter [54].

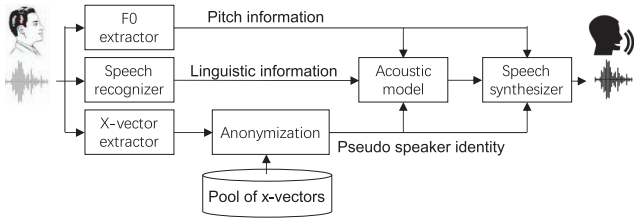



**Fig. 5** Our initial speaker anonymization system including phoneme posteriorgram (PPG) and Mel-spectrogram.

We observed that this system makes it difficult for both machine and human listeners to identify the original speaker from the corresponding anonymized speech. Human listener volunteers compared naturalness between the original and anonymized speech in terms of mean opinion score. No major degradation was observed in the anonymized speech. We also calculated speech quality using the ITU-P.563 Speech Quality Assessment Method. The results demonstrated that the anonymized speech had the same quality as the original speech. However, the linguistic contents were somewhat changed by the anonymization. We subsequently refined this issue by improving the ASR module and increasing the quantity of training data and made the source code publicly available[†].

### 4.3 Gait Anonymization [55]–[58]

Anonymizing gait is as important as face and speech anonymization since gait has become a biometric trait due to its uniqueness for each person and easy recognizability from a distance. The privacy of a person in a video on social media such as YouTube could be compromised by gait recognition systems [59]. Hence, we have proposed gait anonymization methods, which take several different strategies [60]–[62]. The focus of these methods is to deform the silhouettes of the input gait because most modern gait recognition systems use silhouette-based features.

The first strategy we discuss is adding noise to the input gait silhouettes in the feature space [55]. A gait silhouette is considered to consist of a static component, i.e., body

[†]https://github.com/Voice-Privacy-Challenge/Voice-Privacy-Challenge-2020

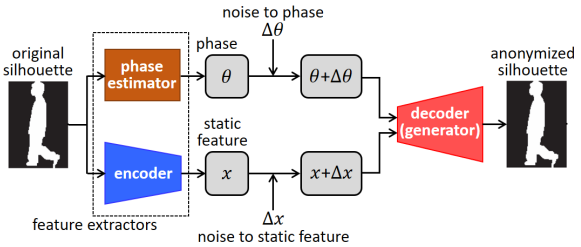



**Fig. 6** Gait anonymization based on noise addition in feature space.

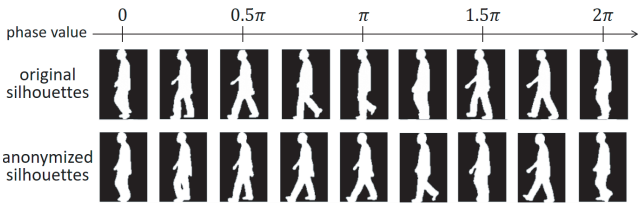



**Fig. 7** Anonymized gait silhouettes generated by adding noise to input gait silhouettes in feature space [55].

shape, and a dynamic component, i.e., posture change. Any posture in the gait cycle is considered to be representable as a phase value ranging from 0 to $2\pi$. The gait silhouettes are anonymized by first decomposing each one into a static feature and a phase value. Next, the feature and value are slightly changed by adding a small noise to each one. Finally, a new silhouette is generated from the changed static feature and phase value, as shown in Fig. 6. The feature extractors and the silhouette generator are trained as an auto-encoder. The addition of noise results in the generated silhouette having no personal identity in both static and dynamic aspects. Moreover, its naturalness is retained because the added noise is small. Figure 7 shows an example of the anonymization results. The success rate of anonymization was 71.1% on average.

The second strategy we discuss is adding a noise directly at the silhouette level. This is done using an auto-encoder network that takes the original gait and a noise gait as inputs [56]. Let $X_1$ and $X_2$ be the original and noise gaits, respectively. The network $\Phi$ is trained by minimizing the loss function

$$\|\Phi(X_1, X_2) - X_1\|^2 + \alpha\|\Phi(X_1, X_2) - X_2\|^2, \quad (1)$$

where the network output $\Phi(X_1, X_2)$ is the anonymization result of $X_1$. The first term is the matching cost between the output and the original gait, and the second term is that between the output and the noise gait. Since it is impossible to simultaneously make these two terms zero, the output $\Phi(X_1, X_2)$ becomes different from the original gait $X_1$ due to the second term. This means the shape of $X_1$ is modified, by which its identity is removed. On the other hand, only small difference is allowed between $\Phi(X_1, X_2)$ and $X_1$ due to the first term, which guarantees the naturalness of $\Phi(X_1, X_2)$. The parameter $\alpha$ specifies how much of the noise gait is added to the original gait. The noise gait is randomly chosen from a pre-constructed dataset so that it differs from the


**Fig. 8** Original and anonymized gaits generated with ST-GAN [57].

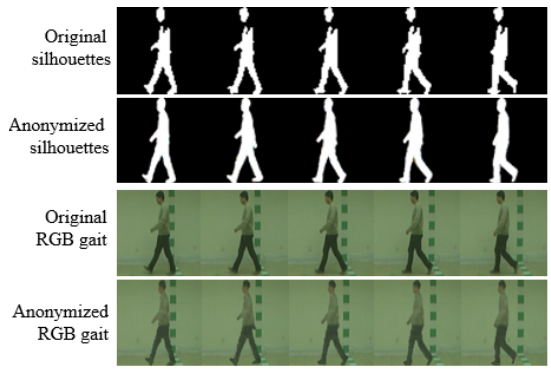



**Fig. 9** Original and anonymized gaits generated from low-quality input silhouettes [58].

original gait and is from the same viewing angle as the original gait. Furthermore, the auto-encoder network described above was extended by using a spatial discriminator, which has a convolutional architecture, and a temporal discriminator, which has a long short-term memory architecture, to generate a more natural gait. This spatio-temporal (ST)-GAN [57] has better anonymization ability than its original model, because it uses random noise synthesized from gait distribution instead of the noise gait. In addition, the ST-GAN can also generate a colorized gait by filling the silhouette-level generation results with the color of the original gait. Several samples generated with the ST-GAN are shown in Fig. 8. The success rate of anonymization ranged from 70.02% to 86.27%.

Methods based on these two strategies can produce natural anonymized gaits when high-quality gait silhouettes are input. However, gait silhouettes are sometimes incorrectly extracted from a video. For instance, several body parts may be missing in the extracted silhouettes. Thus we further proposed a method that can handle low-quality input silhouettes [58]. This method uses a deep convolutional generative adversarial network (DCGAN)-based structure instead of an auto-encoder structure like that of ST-GAN. Since an auto-encoder makes its output as close to its input as possible, low-quality outputs are generated from low-quality inputs. In contrast, a DCGAN-based structure generates a gait in a distribution of the training data. Hence, if it is trained on high-quality silhouettes, it generates natural gaits even from low-quality inputs. Sample outputs are shown in Fig. 9.

### 4.4 Abstraction of User Interest [63], [64]

The interests of cloud-based service users, which can be dis-

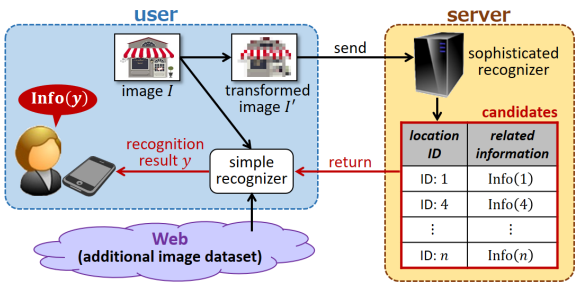



**Fig. 10** Privacy-preserving image recognition framework.

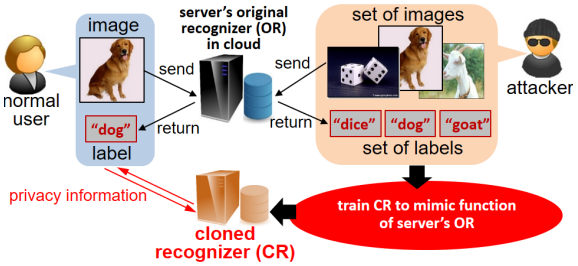



**Fig. 11** *Recognizer cloning attack* on cloud-based services.

cerned from their usage history, are also a type of privacy-sensitive information that should be abstracted. For example, the interests of a user of an image-based information service could be discerned as follows. In a typical image-based information service, a user takes an image (or photograph) of his/her current location with a smartphone and sends it to a cloud server, while the server identifies the location by a sophisticated image recognizer to return the location-related information to the user. Through the use of this service, his/her current location and location history are disclosed to the server in the form of location IDs. To abstract location data and thereby prevent leakage of users' interests, we have proposed a privacy-preserving image recognition framework [63], [64]. An overview is shown in Fig. 10.

As shown in the figure, user image $I$ is transformed to image $I'$ before it is sent to the server. This process is executed on the user's smartphone, which makes it hard for the server to uniquely recognize his/her location from the transformed image. Therefore, the server returns candidate locations rather than the exact one. Then, for each candidate location, the user's smartphone automatically collects typical images of those locations from the Web and compares them with the original image to uniquely determine the exact location $y$, which is the recognition result, as shown in Fig. 10. A simple recognizer can be used for this purpose on the user side because the candidates are preliminarily narrowed down by the server.

In an experiment where the image recognition performance was evaluated under the proposed framework in a real environment, recognition accuracy on the server side was degraded from 99.8% to 41.4% while on the user side it was 86.9%. The proposed framework can thus abstract the interests of image-based information service users.

### 4.5 Prevention of Recognizer Cloning [65]

Image recognizers owned by cloud-based services as mentioned in the previous subsection also should be protected [66]. If a recognizer were to be maliciously stolen or cloned, it could be used by a charge-free fake service, causing economic damage to the original service and a threat to user privacy through leakage of his/her interests. This may become an actual threat in the near future as exemplified by the following scenario. An attacker first sends a large number of unlabeled images to a recognition server and receives the recognition results, i.e., class labels. Next, using the received labels together with the images as a training dataset, the attacker trains a new recognizer that mimics the function of the server's original recognizer (OR). We refer to this attack as a *recognizer cloning attack* (RCA) and refer to a recognizer trained using data obtained in an RCA as a *cloned recognizer* (CR). Figure 11 illustrates an RCA. We proposed two approaches for preventing RCAs [65].

The first approach is to directly prevent the attacker's cloning process. To this end, the server intentionally alters its recognition results so that incorrect labels are sent to the attacker. This leads the attacker to misunderstand the relationships between the images and labels, which degrades the resultant CR. The second approach is to detect trained CRs. We proposed a classifier that receives a pair of recognizers at once and determines whether or not the pair consists of an OR and its CR. If the owner of a cloud-based service uses the classifier on a pair of his/her own OR and any other recognizer, he/she can detect its CRs. Our preliminary experiments revealed that (1) an OR and its CR have almost the same recognition boundary; (2) a CR provides a higher confidence score, which is an output of the recognizer, than its OR [65]. These two characteristics allow us to construct the classifier.

## 5. Modeling of Media Clone Attacks

An essential issue in addressing the media clone problem is the deception caused by confusing real information with fake information. This suggests that, in order to control the risk caused by media clones, we need to understand why such confusion occurs. In addition, fake media will be increasingly used for fraudulent communication such as fraud using advanced technologies of video and speech. Due to the difficulty of technically distinguishing between real and fake media on the basis of their features, it is important to understand what people think when they are deceived.

We focus on people's confusion for existing frauds as a kind of media clone attack. To investigate how people are deceived, we model the communication between the sender or attacker, and the receiver as illustrated in Fig. 12. The attacker tries to commit frauds using phone or email that make the receiver confuse the real with the fake. For this modeling, we introduce channel theory [67], which is a logical formalization to describe how multiple objects

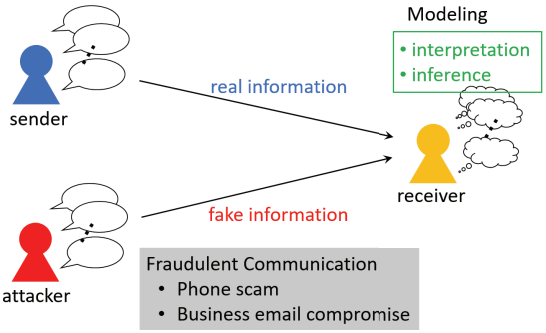



**Fig. 12** Modeling of fraudulent communication.

carry information about each other. The theory was useful to model the human interpretation and inference. For example, Kawakami et al. applied it to analyzing human-human transmission of semantic information with misconception [68], [69]. We claim that channel theory enables us to learn how the receiver's thinking or belief is guided by the attacker.

First, we considered phone scams [70], which occurred frequently over the years in Japan, in order to analyze why the receiver is deceived by the attacker orally pretending to be a relative in distress. We clarified that the receiver's inference with uncertainty can generate his/her new belief that may affect him/her. Second, we investigated business email compromise [71], which occurred in a Japanese airline company, in order to analyze why the receiver is deceived despite having time to investigate the information such as email messages, its attachments and its email addresses. The results indicate that the receiver's inference during email exchange can bring the attacker's logical trap to make him/her confused even if he/she checks the suspicious information. Our approach is capable of objectively tracing the flow of fraudulent communication and modeling the receiver's belief change. This contributes to opening up a new horizon of modeling media clone attacks.

## 6. Conclusion

Media clones made from fake information are a typical example of fake media. This paper has presented a variety of methods for preventing the generation of fake information. The first step in preventing MC attacks is to protect the information of a specific person so that fake information cannot be created from his/her authentic information. The methods we have developed in the MC project contribute to this effort. However, it is impossible to protect all the media information of celebrities such as politicians and TV personalities because their face or voice is repeatedly delivered via TV, YouTube, and other sources every day. Media clones of such celebrities could be created from a collection of their media information. Therefore, the next step is to detect media clones on the end-user (receiver) side, which is of great importance. For this purpose, diverse approaches to generating elaborate media clones should be investigated. Detection and generation of media clones will be discussed in a subsequent paper [2].

## References

[1] N. Babaguchi, "Communication system for defending against attacks of media clones." https://www.jsps.go.jp/j-grantsinaid/12_kiban/ichiran_28/e-data/h28_eng_16h06302.pdf, accessed March 23. 2020.
[2] I. Echizen, N. Babaguchi, J. Yamagishi, N. Nitta, Y. Nakashima, K. Nakamura, K. Kono, F. Fang, S. Myojin, Z. Kuang, H.H. Nguyen, and N.D.T. Tieu, "Generation and detection of media clones," IEICE Trans. Inf. & Syst., vol.E104-D, no.1, pp.12–23, Jan. 2021.
[3] A. Harvey, "Camouflage from face detection." https://cvdazzle.com/, accessed April 6. 2020.
[4] T. Yamada, S. Gohshi, and I. Echizen, "Use of invisible noise signals to prevent privacy invasion through face recognition from camera images," Proc. 20th ACM Multimedia Conf., pp.1315–1316, 2012.
[5] T. Yamada, S. Gohshi, and I. Echizen, "Privacy Visor: Method for preventing face image detection by using differences in human and device sensitivity," Proc. Joint IFIP TC6 and TC11 Conf. on Communications and Multimedia Security, pp.152–161, 2013.
[6] M. Sharif, S. Bhagavatula, L. Bauer, and M. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," Proc. ACM SIGSAC Conf. on Computer and Communications Security, pp.1528–1540, 2016.
[7] K. Hashimoto, J. Yamagishi, and I. Echizen, "Privacy-preserving sound to degrade automatic speaker verification performance," Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp.5500–5504, 2016.
[8] S. Thys, W.V. Ranst, and T. Goedeme, "Fooling automated surveillance cameras: Adversarial patches to attack person detection," Proc. IEEE Conf. on Computer Vision Workshop, 7 pages, 2019.
[9] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.Y. Chen, Y. Wang, and X. Lin, "Adversarial t-shirt! Evading person detectors in a physical world," arXiv:1910.11099, 14 pages, 2019.
[10] V. Ayala-Rivera, P. McDonagh, T. Cerqueus, and L. Murphy, "A systematic comparison and evaluation of k-anonymization algorithms for practitioners," Trans. Data Privacy, vol.7, no.3, pp.337–370, 2014.
[11] B. Bhattarai, A. Mignon, F. Jurie, and T. Furon, "Puzzling face verification algorithms for privacy protection," Proc. IEEE Int. Workshop on Information Forensics and Security, pp.66–71, 2014.
[12] S. Ribaric and N. Pavesic, "An overview of face de-identification in still images and videos," Proc. IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition, pp.1–6, 2015.
[13] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S.K. Nayar, "Face swapping: Automatically replacing faces in photographs," ACM Trans. Graphics, vol.27, no.3, pp.39:1–39:8, 2008.
[14] P. Korshunov and T. Ebrahimi, "Using face morphing to protect privacy," Proc. IEEE Int. Conf. on Advanced Video and Signal Based Surveillance, pp.208–213, 2013.
[15] L. Yuan, P. Korshunov, and T. Ebrahimi, "Privacy-preserving photo sharing based on a secure JPEG," Proc. IEEE Conf. on Computer Communications Workshops, pp.185–190, 2015.
[16] S.J. Oh, R. Benenson, M. Fritz, and B. Schiele, "Faceless person recognition: Privacy implications in social media," Proc. European Conf. on Computer Vision, pp.19–35, 2016.
[17] K. Brkic, I. Sikiric, T. Hrkac, and Z. Kalafatic, "I know that person: Generative full body and face de-identification of people in images," Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop, pp.1319–1328, 2017.
[18] N. Ruchaud and J.-L. Dugelay, "Automatic face anonymization in visual data: Are we really well protected?," Proc. Electronic Imaging, Image Processing: Algorithms and Systems XIV, pp.1–7, 2016.
[19] Y. Nakashima, T. Ikeno, and N. Babaguchi, "Evaluating protection

capability for visual privacy information," IEEE Security Privacy, vol.14, no.1, pp.55–61, 2016.
[20] L. Sweeney, "k-anonymity: A model for protecting privacy," Int. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol.10, no.05, pp.557–570, 2002.
[21] E.M. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying face images," IEEE Trans. Knowl. Data Eng., vol.17, no.2, pp.232–243, 2005.
[22] S. Ribaric, A. Ariyaeeinia, and N. Pavesic, "De-identification for privacy protection in multimedia content: A survey," Signal Processing: Image Communication, vol.47, pp.131–151, 2016.
[23] R. Gross, L. Sweeney, F. De la Torre, and S. Baker, "Model-based face de-identification," Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop, p.161, 2006.
[24] Z. Sun, L. Meng, and A. Ariyaeeinia, "Distinguishable de-identified faces," Proc. IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition, pp.1–6, 2015.
[25] B. Meden, Ž. Emeršič, V. Štruc, and P. Peer, "k-Same-Net: k-Anonymity with generative deep neural networks for face deidentification," Entropy, vol.20, no.1, pp.60:1–60:24, 2018.
[26] Z. Ren, Y. Jae Lee, and M.S. Ryoo, "Learning to anonymize faces for privacy preserving action detection," Proc. European Conf. on Computer Vision, pp.620–636, 2018.
[27] Y. Wu, F. Yang, and H. Ling, "Privacy-protective-GAN for face de-identification," arXiv:1806.08906, 11 pages, 2018.
[28] O. Gafni, L. Wolf, and Y. Taigman, "Live face de-identification in video," Proc. Int. Conf. on Computer Vision, pp.9377–9386, 2019.
[29] H. Hukkelås, R. Mester, and F. Lindseth, "DeepPrivacy: A generative adversarial network for face anonymization," Proc. Int. Symposium on Visual Computing, pp.565–578, 2019.
[30] T. Li and L. Lin, "Anonymousnet: Natural face de-identification with measurable privacy," Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 10 pages, 2019.
[31] M. Pobar and I. Ipšić, "Online speaker de-identification using voice transformation," Proc. Int. Convention on Information and Communication Technology, Electronics and Microelectronics, pp.1264–1267, 2014.
[32] T. Justin, V. Štruc, S. Dobrišek, B. Vesnicer, I. Ipšić, and F. Mihelič, "Speaker de-identification using diphone recognition and speech synthesis," Proc. IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition, pp.1–7, 2015.
[33] C. Magariños, P. Lopez-Otero, L. Docio-Fernandez, E.R. Banga, C. Garcia-Mateo, and D. Erro, "Piecewise linear definition of transformation functions for speaker de-identification," Proc. Int. Workshop on Sensing, Processing and Learning for Intelligent Machines, pp.1–5, 2016.
[34] Q. Jin, A.R. Toth, T. Schultz, and A.W. Black, "Speaker de-identification via voice transformation," Proc. Workshop on Automatic Speech Recognition Understanding, pp.529–533, 2009.
[35] F. Bahmaninezhad, C. Zhang, and J. Hansen, "Convolutional neural network based speaker de-identification," Proc. Odyssey 2018: The Speaker and Language Recognition Workshop, pp.255–260, 2018.
[36] C. Magariños, P. Lopez-Otero, L. Docio-Fernandez, E. Rodriguez-Banga, D. Erro, and C. Garcia-Mateo, "Reversible speaker de-identification using pre-trained transformation functions," Computer Speech & Language, vol.46, pp.36–52, 2017.
[37] J. Han and B. Bhanu, "Statistical feature fusion for gait-based human recognition," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.842–847, 2004.
[38] M.A. Hossain, Y. Makihara, J. Wang, and Y. Yagi, "Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control," Pattern Recognition, vol.43, no.6, pp.2281–2291, 2010.
[39] A. Tsuji, Y. Makihara, and Y. Yagi, "Silhouette transformation based on walking speed for gait identification," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp.717–722, 2010.
[40] P. Agrawal and P. Narayanan, "Person de-identification in videos," IEEE Trans. Circuits Syst. Video Technol., vol.21, no.3, pp.299–310, 2011.
[41] G. Momose and H. Takahashi, "Gait feature analysis for personal de-identification," Proc. NICOGRAPH International, pp.130–133, 2012.
[42] T. Ogane and I. Echizen, "BiometricJammer: Use of pseudo fingerprint to prevent fingerprint extraction from camera images without inconveniencing users," Proc. 2018 IEEE Int. Conf. on Systems, Man, and Cybernetics, pp.2825–2831, 2018.
[43] T. Ogane and I. Echizen, "BiometricJammer: Preventing surreptitious fingerprint photography without inconveniencing users," Proc. Int. Joint Conf. on Biometrics, pp.253–260, 2017.
[44] B.C. Club, "Fingerprint biometrics hacked again." https://www.ccc.de/en/updates/2014/ursel, Accessed Feb. 27, 2020.
[45] F. Fang, J. Yamagishi, I. Echizen, M. Sahidullah, and T. Kinnunen, "Transforming acoustic characteristics to deceive playback spoofing countermeasures of speaker verification systems," Proc. 2018 IEEE Int. Workshop on Information Forensics and Security, pp.1–9, 2018.
[46] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," Proc. Interspeech, pp.3642–3646, 2017.
[47] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K.A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," Proc. Annual Conf. of the Int. Speech Communication Association, 5 pages, Aug. 2017.
[48] T. Kinnunen, K.A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D.A. Reynolds, "t-DCF: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification," arXiv:1804.09618, 8 pages, 2018.
[49] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with GANs," Int. Journal of Computer Vision, 16 pages, 2019.
[50] Y. Nakashima, T. Yasui, L. Nguyen, and N. Babaguchi, "Speech-driven face reenactment for a video sequence," ITE Trans. Media Technology and Applications, vol.8, no.1, pp.60–68, 2020.
[51] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, "Speaker anonymization using x-vector and neural waveform models," Proc. 10th ISCA Speech Synthesis Workshop, pp.155–160, 2019.
[52] B. BabaAli and K. Vesely, "Kaldi timit recipe." https://github.com/kaldi-asr/kaldi/blob/master/egs/timit, accessed March 23. 2020.
[53] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," Proc. 2018 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp.5329–5333, 2018.
[54] X. Wang, S. Takaki, and J. Yamagishi, "Neural source-filter-based waveform model for statistical parametric speech synthesis," Proc. 2019 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp.5329–5333, 2019.
[55] Y. Hirose, K. Nakamura, N. Nitta, and N. Babaguchi, "Anonymization of gait silhouette video by perturbing its phase and shape components," Proc. Asia-Pacific Signal and Processing Association Annual Summit and Conf., pp.1679–1685, 2019.
[56] N.-D.T. Tieu, H.H. Nguyen, H.-Q. Nguyen-Son, J. Yamagishi, and I. Echizen, "An approach for gait anonymization using deep learning," Proc. 2017 IEEE Workshop on Information Forensics and Security, pp.1–6, 2017.
[57] N.-D.T. Tieu, H.H. Nguyen, H.-Q. Nguyen-Son, I. Echizen, and J. Yamagishi, "Spatio-temporal generative adversarial network for gait anonymization," Journal of Information Security and Applications, vol.46, pp.307–319, June 2019.
[58] N.-D.T. Tieu, H.H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "An RGB gait anonymization model for low-quality silhouettes," Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conf. 2019, pp.1686–1693, Nov. 2019.

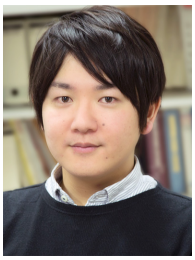
[59] Y. Makihara, D.S. Matovski, M.S. Nixon, J.N. Carter, and Y. Yagi, "Gait recognition: Databases, representations, and applications," John Wiley & Sons, Inc., pp.1–15, 2015.
[60] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," Proc. Int. Conf. on Biometrics, pp.1–8, 2016.
[61] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," Proc. Crime Detection and Prevention, pp.1–6, 2009.
[62] J. Han and B. Bhanu, "Individual recognition using gait energy image," IEEE Trans. Pattern Anal. Mach. Intell., vol.28, no.2, pp.316–322, 2006.
[63] K. Nakamura, N. Nitta, and N. Babaguchi, "Encryption-free framework of privacy-preserving image recognition for photo-based information services," IEEE Trans. Inf. Forensics Security, vol.14, no.5, pp.1264–1279, 2019.
[64] K. Fujii, K. Nakamura, N. Nitta, and N. Babaguchi, "A framework of privacy-preserving image recognition for image-based information services," Proc. Int. Conf. on Multimedia Modeling, pp.40–52, 2017.
[65] K. Nakamura, N. Nitta, and N. Babaguchi, "Investigation of methods for defending against recognizer clones," Medical Imaging Technology, vol.37, no.4, pp.188–193, 2019. (in Japanese).
[66] F. Tramer, F. Zhang, A. Juels, M.K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction apis," Proc. USENIX Conf. on Security Symposium, pp.601–618, 2016.
[67] J. Barwise and J. Seligman, Information flow -the logic of distributed systems, Cambridge University Press, 1997.
[68] H. Kawakami, H. Suto, H. Handa, O. Katai, and T. Shiose, "Analyzing diverse interpretation as benefit of inconvenience," Proc. Int. Symposium on Symbiotic Nuclear Power Systems for 21st Century, pp.75–81, 2009.
[69] H. Kawakami, K. Nishitani, T. Shiose, and O. Katai, "Mathmatical analysis for benefit of diverse interpretation," Proc. SICE Symposium on Intelligent Systems, 6 pages, 2008. (in Japanese).
[70] S. Myojin and N. Babaguchi, "A logical consideration on deceived person's thinking," Artificial Life and Robotics, vol.24, no.1, pp.114–118, March 2019.
[71] S. Myojin and N. Babaguchi, "A logical consideration on fraudulent email communication," Artificial Life and Robotics, vol.25, no.3, pp.475–481, Aug. 2020.

**Noboru Babaguchi** received B.E., M.E., and Ph.D. degrees in communication engineering from Osaka University, Japan, in 1979, 1981, and 1984, respectively. He is currently a professor and the dean of the Graduate School of Engineering, Osaka University. His research interests include image/video analysis, multimedia computing, and intelligent systems. He has recently been researching privacy protection methods for visual information and the security and fabrication of multimedia. He has published over 250 journal and conference papers and several textbooks. He served as a guest editor of the IEEE Transactions on Information Forensics and Security, Special Issue on Intelligent Video Surveillance for Public Security & Personal Privacy. He also served as a general co-chair of MMM2008, a general co-chair of ACM Multimedia 2012, a track co-chair of ICPR2012, an area chair of IEEE ICME2013, and an honorary co-chair of ACM ICMR2018. He is a fellow of the IEICE, a vice president of the ITE, a senior member of the IEEE, and a member of the ACM, IPSJ, and JSAI.

**Isao Echizen** received B.S., M.S., and D.E. degrees from the Tokyo Institute of Technology, Japan, in 1995, 1997, and 2003, respectively. He joined Hitachi, Ltd. in 1997 and until 2007 was a research engineer in the company's systems development laboratory. He is currently an advisor to the general of the National Institute of Informatics (NII), a professor in NII's Information and Society Research Division, and a professor in the Department of Information and Communication Engineering, Graduate School of Information Science and Technology, The University of Tokyo, Japan. He is also a visiting professor at the Tsuda University, Japan, and was a visiting professor at the University of Freiburg, Germany, in 2010 and at the University of Halle-Wittenberg, Germany, in 2011. He is currently engaged in research on information security and content security and privacy. He received the Best Paper Award from the IPSJ in 2005 and 2014, the Fujio Frontier Award and the Image Electronics Technology Award in 2010, the One of the Best Papers Award from the Information Security and Privacy Conference in 2011, the IPSJ Nagao Special Researcher Award in 2011, the DOCOMO Mobile Science Award in 2014, the Information Security Cultural Award in 2016, and the IEEE Workshop on Information Forensics and Security Best Paper Award in 2017. He was a member of the Information Forensics and Security Technical Committee and the IEEE Signal Processing Society. He is the Japanese representative on IFIP TC11 (Security and Privacy Protection in Information Processing Systems).

**Junichi Yamagishi** received a Ph.D. by Tokyo Institute of Technology in 2006. He was a senior research fellow in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh, U.K., from 2006 to 2013. He is currently a professor at National Institute of Informatics in Japan. He was awarded the Itakura Prize from the Acoustic Society of Japan, the Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan, and the Young Scientists' Prize from the Minister of Education, Science and Technology, the JSPS prize, and DOCOMO prize in 2010, 2013, 2014, 2016, and 2018, respectively. He served previously as co-organizer for the bi-annual ASVspoof special sessions at INTERSPEECH 2013-9, the bi-annual Voice conversion challenge at INTERSPEECH 2016 and Odyssey 2018, an organizing committee member for the 10th ISCA Speech Synthesis Workshop 2019 and a technical program committee member for IEEE ASRU 2019. He also served as a member of the IEEE Speech and Language Technical Committee, as an Associate Editor of the IEEE/ACM TASLP and a Lead Guest Editor for the IEEE JSTSP SI on Spoofing and Countermeasures for Automatic Speaker Verification. He is currently a PI of JST-CREST and ANR supported VoicePersona project. He also serves as a chairperson of ISCA SynSIG and as a Senior Area Editor of the IEEE/ACM TASLP.

**Naoko Nitta** received the B.E., M.E., and Ph.D. degrees in Engineering from Osaka University, in 1998, 2000, and 2003, respectively. She is currently an Associate Professor in Graduate School of Engineering, Osaka University. From 2002 to 2004, she was a research fellow of the Japan Society for the Promotion of Science. From 2003 to 2004, she was a Visiting Scholar at Columbia University. Her research interests are in the areas of multimedia content and social media analysis.



**Yuta Nakashima** received the B.E. and M.E. degrees in communication engineering and the Ph.D. degree in engineering from Osaka University, Osaka, Japan, in 2006, 2008, and 2012, respectively. From 2012 to 2016, he was an Assistant Professor at the Nara Institute of Science and Technology. He is currently an Associate Professor at the Institute for Datability Science, Osaka University. His research interests include computer vision and machine learning and their applications.

**Kazuaki Nakamura** received the B.S. degree in Engineering from Kyoto University in 2005, and the M.S. and Ph.D. degrees in Informatics from Kyoto University in 2007 and 2011, respectively. He is currently an Assistant Professor at Graduate School of Engineering, Osaka University, from 2012. His research interests include image processing, image recognition, and video analysis. He is a member of IEEE, IEICE, IPSJ, and ITE.

**Kazuhiro Kono** received the B.E, M.E., and Ph.D. degrees in communication engineering from Osaka University, Japan, in 2005, 2007, and 2010, respectively. He is currently an Associate Professor in the Faculty of Societal Safety Sciences, Kansai University. His research interests include information security and privacy or personal information technologies. He is a member of IEICE, IPSJ, REAJ, IEEE, and ACM.

**Fuming Fang** received B.S. degree in automation mechanic from Changchun University, Changchun, China, in 2008. He received M.S. degree in informatics from Chiba University, Chiba, Japan, in 2013, and received Ph.D. degree in informatics in Tokyo Institute of Technology, Tokyo, Japan, in 2017. After that, he joined National Institute of Informatics as a project researcher. His research interests include information security, machine learning, voice conversion, speech synthesis, speaker recognition, speech recognition, and neural language processing.

**Seiko Myojin** received a Ph.D. degree in engineering from Osaka University, Japan. She is currently a specially appointed assistant professor in the Graduate School of Engineering, Osaka University, Japan. Her current research interests include phenomena caused by media surrounding human. She is qualified as a senior virtual reality specialist (VRSJ). She is currently a member of the IPSJ, IEICE, VRSJ, SICE, and HIS.

**Zhenzhong Kuang** received M.S. and Ph.D. degrees from China University of Petroleum, Qingdao, China in 2013 and 2017, respectively. From 2015 to 2017, he was with University of North Carolina at Charlotte, Charlotte, USA. From 2018 to 2019, he was a Researcher with the Media Integrated Communication Lab., Graduate School of Engineering, Osaka University, Osaka, Japan. He currently works in with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou, China. His research interests include image privacy protection, multimedia analysis and machine learning.

**Huy H. Nguyen** received B.S. degree in Information Technology from VNUHCM - University of Science, Ho Chi Minh City, Vietnam in 2013. He is currently pursuing a Ph.D. degree in computer science at the Graduate University for Advanced Studies (SOKENDAI), Kanagawa, Japan. Her current research interests include security and privacy in biometrics and machine learning.

**Ngoc-Dung T. Tieu** received B.S. degree in Information Technology from Hanoi University of Science and Technology, Hanoi, Vietnam in 2003 and M.S. in Electronics and Computer Engineering from Korea University, Seoul, Korea in 2006. During 2008-2016, she was a lecturer at University of Transport and Communications, Hanoi, Vietnam. She is currently pursuing a Ph.D. degree in computer science at the Graduate University for Advanced Studies (SOKENDAI), Kanagawa, Japan. Her current research interests include information security, machine learning and image processing.