A Taxonomy of Audiovisual Fake Multimedia Content Creation Technology

Ali Khodabakhsh, Raghavendra Ramachandra, Christoph Busch
Norwegian Biometrics Laboratory
Norwegian University of Science and Technology
Gjøvik, Norway
{*ali.khodabakhsh,raghavendra.ramachandra,christoph.busch*}*@ntnu.no*

***Abstract*—The spread of fake and misleading multimedia content on social media has become commonplace and is effecting society and its decision procedures negatively in many ways. One special case of exploiting fake content is where the deceiver uses the credibility of a trustworthy source as the means of spreading disinformation. Thanks to advancements in technology, the creation of such content is becoming possible in audiovisual form with limited technical knowledge and at low cost. The potential harm of circulation of these content in media calls for the development of automated detection methods. This paper offers a categorization of such fake content creation technology in an attempt to facilitate further study on generalized countermeasures for their detection.**

***Keywords*-Multimedia forensics; Fake content detection; Fake multimedia content creation technology; Content manipulation detection**

I. Introduction

Consumption of digital media and its impact on decision procedures (e.g. elections) has reached a majority-owned relevance over traditional media (e.g printed newspapers) in our world of ubiquitous information devices (Smartphone, tablets). Along with that cultural change, we must accept for the consumed content an inherent loss of data authenticity. The lack of proper fact-checking and third-party filtering on these platforms compared to traditional media resulted in the prevalence of misinformation and disinformation on these media [1]. The spread of fake content can have a long-lasting impact on individuals opinions even after presentation of factual information [2].

One special case of fake content is where a deceiver uses the identity of another person (e.g. an authority figure) to disseminate false information, taking advantage of his/her credibility. Recent advancements in technology made it possible to create such content in audiovisual form (Fig. 2i) [3], [4], using commodity devices, and at low cost. A demonstration of existing technologies has been made available online for the purpose of public awareness: http://futureoffakenews.com.

These content are of special importance as talking faces are a natural way of communication for humans, and are preferred to other forms of communication. Furthermore, despite considerable progress on detection of fake textual content [5], very little effort has been directed to protect consumers from fake multimedia content. On the other hand, manual detection is very costly and the capacity of authentication can be out-competed by the mass of user-generated content. "*Personation*" is defined by the *Oxford English Dictionary* as "*The action of assuming a character, or of passing oneself off as someone else, esp. for fraudulent purposes*"[1]. In the context of this study, audiovisual personation can be described as any attempt to assume the identity of another person in a audiovisual form, with intent to deceive. The cases of convincing personations in history have been limited to people with natural similarity (e.g. the actor Clifton James, who resembled General Montgomery in a deception mission in World War II [6]). However, as technology advances, a wide range of virtual and artificial personation techniques are becoming available, and examples of their use can be found in many real-life applications.

For personations to be successful in deception, the created content should be of high quality to pass the multimodal judgment of naive media consumers in naturalness and similarity of speech, appearance, and behavior. As a result, they should be based on a good understanding of the human perception of reality and identity. A notably related concept is the uncanny valley [7] hypothesis. This hypothesis states that after a specific point, the more an artificial entity resembles a human outlook and behavior, the presentation will elicit a more negative emotional response from the observer. Nevertheless, despite the difficulty and expenses of climbing up again from the depth of the uncanny valley, a vast amount of effort has been dedicated to the creation of realistic artificial humans, and many instances of artificial entities have achieved realism in the sense of being indistinguishable from reality to unsuspecting humans. The Hollywood industry with its need for realistic yet low-cost animated scenes stimulated significant innovation in this domain over recent years.

This article proposes categories to group existing technologies for the creation of plausible audiovisual personation content with the goal of providing a comprehensive overview of deception attempts and creating a ground for the development of generalized detectors.

The rest of this paper is organized as follows: Section II

[1]"personation, n." OED Online. Oxford University Press, June 2017. Web. 21 December 2017.

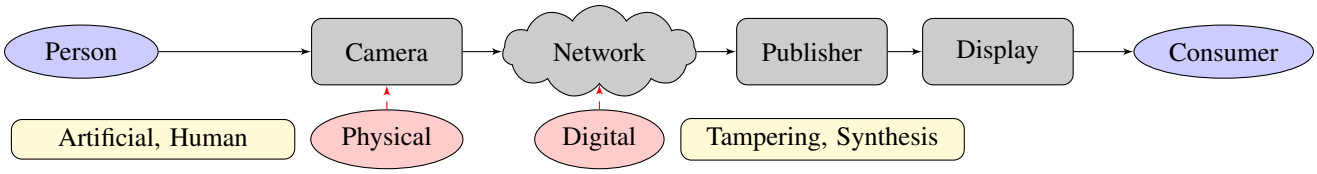


Figure 1: Points of vulnerability of video transmission medium to deception attempts.

describes the technologies and the motivation behind the development of these technologies. Section III describes briefly the existing detection methods. Section IV summaries the study and discusses its implications, and finally section V will conclude and describe future work.

II. Personation Methods

The technology for generation of artificial lifelike human appearance is advancing with the goal of creating the experience of submersion and a greater degree of presence and natural interaction with the artificial entity. The consumer may be aware of the unreality of the entity, however, the apparent realism makes it cognitively possible to have suspension of disbelief. The artificial entities may be digital (e.g. an avatar), or physical (e.g. an android robot). These technologies have applications in communication (e.g. telepresence, customer service, advertisement) , training (e.g. education, simulation) , health-care (elderly care, physical and psychological therapy) , assistance (companionship, museum guides, office robots, software office assistants) , entertainment (e.g. cinematography, satire, video games, stage shows) , and covert disinformation attacks. Based on the application, the resulting systems can create a passive representation, or be interactive.

For the purpose of this article, the technologies can be categorized by the point of application in the consumption process of the audiovisual content. This is motivated by the difference in technical demands of content generation, and thus detection approaches at each application point. Fig. 1 shows the lifetime of an audiovisual content. A video depicting a person is recorded by a camera, and after traveling the network (including storage devices), it is shared by a publisher and displayed to the consumer. Given an audiovisual representation of a person, the points of suspicion are a false presentation at the camera, digital tampering of the recorded video, or replacement with a computer generated (CG) counterpart.

Based on these points of vulnerability and different modalities of the audiovisual signal, the following categorization is used to cover existing personation technology which will be discussed first for visual and subsequently for audio content:

A. Visual

A visual personation requires naturalness and similarity to the target person in appearance and behavior. The behavior of the personation can be modeled and applied independently of the appearance, and thus it is described separately.

1) Physical: A physical visual personation requires a convincing appearance of a person or an object with the resemblance of the target person. This item can be created as an artist's impression, or be created using the scan or cast of the face of a person.

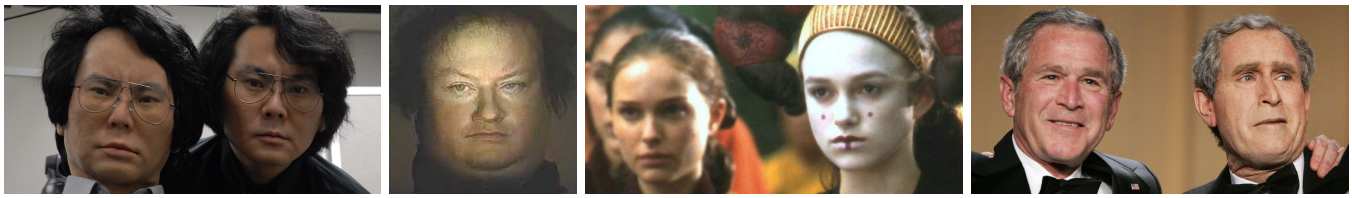
Artificial: Artificial visual personation can be described as any physical artifact (i.e. movable dummies and fleshly robots) that can convincingly resemble the target person in appearance and ability to move. Due to the complexity of the human facial muscle configuration and movements, it is not possible to puppeteer the artifact mechanically. Thus the artificial personation devices are usually operated by robots. Such robots are called androids and can have a photorealistic resemblance to the target person thanks to realistic skin and hair like material used in their production. These androids are mainly developed by robotics community for natural human-robot interactions, and have applications ranging from entertainment to education and health-care. Notable examples are animatronics of US presidents at the hall of presidents in the Walt Disney world resort[2], and Geminoid robots (Fig. 2a) created by Hiroshi Ishiguro at the Intelligent Robotics Lab at Osaka University [8].

The facial movements are typically modeled by motors acting facial action units on the face. Due to mechanical limitations, these robots have jerky movements and their behavior is easily detectable as unnatural. To avoid these limitations in facial motion, some androids use a screen as a face (e.g. Life Imaging Projection System aka L.I.P.S (Fig. 2b))[3]. Another notable example is the shape-shifting robot WD-2, which can replicate the face of a person based on the 3D scan of his face. The high cost of building and the unnatural movements limits the application of these androids in personation attempts.

Human: This category is the oldest personation method that has been used for deception. The cost of personation varies depending on the apparent natural similarity (i.e. biometric twins) of the target person and the personator. In case of lack of sufficient resemblance, the personator can use heavy or prosthetic makeup and masks to change his

[2]http://www.popularmechanics.com/technology/robots/a23699/robot-presidents-disney/

[3]https://news.yale.edu/2001/03/19/heads-will-be-talking-yales-digital-media-arts-center



(a) The Geminoid [8] (left) and Hiroshi Ishiguro (right)

(b) Life Imaging Projection System[3]

(c) Natalie Portman (left) and Keira Knightley (right) [9]

(d) George W. Bush (left) and Steve Bridges (right) [10]

(e) Ezzat et al. [11]

(f) Aimi Eguchi[4]

(g) Keanu Reeves (left) and his digital double (right) [12]

(h) Seymour et al. [13]

(i) Thies et al. [4]

(j) Suwajanakorn et al. [14]

Figure 2: Illustration of different visual personation technologies.

appearance[5]. The result is often of sufficient similarity to be recognized as the target person. Many applications of this technique exist and are mainly around the entertainment industry, such as "fake shemps"[6] and impersonators.

An example for identical twins is Leslie H. Gearren[7] acting for Linda Hamilton in "*Terminator 2: Judgment Day*" as her double. Natural similarity of actors Keira Knightley for Natalie Portmans character has also been used in "*Star Wars: Episode I The Phantom Menace*" (Fig. 2c). Many examples for prosthetic makeup exist in satire (e.g. Steve Bridges as George Bush (Fig. 2d) [10]). Using humans for personation has been done for political purposes too. The best-documented example of political decoys is personation of Bernard Montgomery by Clifton James [6]. This method of personation is surprisingly effective in convincing people. The main advantage of this technique compared to the other methods is complete naturalness of the muscle control of the resulting personation.

The impersonator needs to learn the gestures and mannerism of the target person in order for the personation to be convincing. For such applications, actors are usually the best choice as of their experience in realistic mimicking of behavior. This will provide similarity on top of realism of their movement.

2) Digital: Using computer algorithms, a video of a talking face can be a digitally modified copy or be completely synthetic. Different technologies evolved for the creation of animated faces based on these two categories for applications such as virtual actors and automated dubbing.

Tampering: An authentic video of a person can be manipulated and modified to change the content of the recording. This can be done manually using video editing software (Fig. 2f) (e.g. splicing and morph cut in Adobe Premiere) or automatically using techniques such as active appearance models (AAM) [11]. These changes can require signal processing steps minimally, as of removing a single word manually and morphing the before and after images, or extensively, as for automatic concatenation of visemes in audiovisual text-to-speech (AVTTS).

One of the earliest examples of automatic tampering is Video Rewrite system [15]. Since these methods produce the original frames of the recorded video or their morphed copy, the result is generally photorealistic and similar to the target person. However, the realism of dynamics is limited by the amount of variability in the existing footage. The more tampering and morphing happens between incoherent frames, the more temporal artifacts will be visible in the resulting video. This method has been successfully used for AVTTS and achieved high realism scores in subjective tests (Fig. 2e) [11]. The limitation of this method is that it requires a long expressive recording with consistent light and a fixed pose for desirable results. However, the capture and animation process is much simpler and computationally cheaper compared to synthesis and results in higher quality videos.

Synthesis: The high computational cost and difficulty in 3D modeling of human facial details and rendering of digital characters, as well as the extreme sensitivity of humans to details of facial texture and motion, makes the generation of synthetic faces hard. However, due to the flexibility these models provide for synthesis in different lighting conditions, from different angles, and with the minimal amount of

[4]http://newsfeed.time.com/2011/06/24/japanese-scientists-build-a-perfect-and-fake-pop-star/

[5]https://www.boredpanda.com/game-of-thrones-make-up-art-transformation-paolo-ballesteros/

[6]https://web.archive.org/web/20071115162315/http://en.allexperts.com/q/Horror-Film-2863/Horror-Film-Staff.htm

[7]http://www.imdb.com/name/nm0357696/bio

capture needed compared to tampering techniques, there has been a lot of interest and effort in creating realistic synthetic faces [16]. The existing technology has been used to synthesize faces of sufficient realism by the movie industry in the past decade (Fig. 2g). However, the realism of synthesis is a function of computational costs such as the number of polygons and reflection and shading resolution, making the technology limited to high budget non-realtime applications. Nevertheless, in some cases, it may be possible to reduce computational costs by only synthesizing the face partially and splicing it over some existing footage [16].

The advancements in computational graphics and graphics processing unit capacity slowly bring the possibility of photorealistic 3D rendering to real-time and on personal computers (Fig. 2h) [13]. The capture procedure of faces usually requires the use of multiview face capture systems [17]. It has also recently become possible to infer the high-resolution texture of faces using a single low-resolution photo of the face [18]. Morph target animation can be used along with facial rigging to animate the face mesh.

These models can present very high photorealism [13] thanks to methods for perfecting the details (e.g. skin reflectance modeling) [17]. These synthetic faces have also found applications in AVTTS [11] and robotics [19].

3) Animation source: The aforementioned physical and digital artificial entities have interfaces for animation (e.g. based on FACS). To answer how to animate these characters using their interface, there are several solutions developed and are described in this section.

Motion capture: Motion capture (mocap) technology has advanced tremendously recently, and many markerless mocap systems have been developed with high accuracy [20]. This enables the actor or impersonator to control the actions of the virtual or artificial character with ease and accuracy. Based on the resolution of the motion capture device, the movements can be indistinguishable from real movements. These systems have been applied by the movie industry as well as for virtual reality and telepresence applications.

Synthesis: In many cases, it is not possible to entirely rely on motion capture for animation of the characters. Examples include video games and autonomous robots. Early systems were animated using predefined actions that were coded manually [21]. Example of these systems are the terminal-analog systems that were early attempts to animate AVTTS characters. There have been attempts to animate characters automatically using models such as hidden Markov models (HMMs). These models can be trained on existing footage of the target person, and used for the synthesis of proper behavior in new situations. Another type of synthesis is the use of text or speech features to animate the character in the video (Fig. 2j) [14]. These systems have applications in AVTTS as well as automated dubbing.

B. Auditory

Humans rely on dynamics and high-level auditory features for recognizing people, and vocal-tract similarity does not affect the human perception as much as the dynamics of speech. The resulting situation requires realistic virtual and physical artificial beings to have natural sounding personations, as well as having similarity in high-level features.

1) Physical: Physical methods rely on physical entities for generation of personation speech. These can be broadly categorized into artificial and living.

Artificial: A speech personation audio can be generated using biomechanical modeling of human vocal apparatus [22]. These systems are hard to develop as the vocal apparatus of humans is not visible and not measurable as easily as faces. The limitations are similar to those of artificial visual personation technologies. The technology has not reached maturity for use in personation.

Human: Professional impressionists can successfully imitate the voice of many different people. This ability shows that no alteration to the vocal tract is needed, and impersonation is an ability that can be learned by practice. Impersonations usually mimic the mannerism of the target person and try to adjust their voice dynamics to match that of him/her. The resulting speech is convincingly similar and sounds natural to the human ear. Impersonation is usually used by impressionists for entertainment, however, instances of their use have been recorded for personal and political gains. A notable example is the personation of President Truman's voice on the telephone to persuade foreign leaders to vote in particular ways at the United Nations[8].

2) Digital: Speech signal can be manipulated and generated digitally as well. Many different systems have been developed with high naturalness and intelligibility for real-life applications. Similar to digital visual personation techniques, these techniques are also categorizable to tampering and synthesis.

Tampering: A synthetic speech can be generated by concatenation of speech samples from a target speaker. The concatenation footprints can be minimal, in the case of removal of a word from an audio, or audible when extensively done (e.g. diphone synthesis). The automated systems generating this kind of synthetic speech are typically called unit-selection speech synthesis systems [23]. Due to the use of natural human voice for the generation of the synthetic speech, the resulting audio has very natural human-like sound, resembling the voice of the target speaker. However, due to the collection of each unit from a different context, the high-level features such as style and intonation of speech are often lost. Some of these artifacts can be corrected using post-processing of the pitch and duration of phonemes after synthesis (e.g. using Pitch Synchronous Overlap and Add (PSOLA)). Unit-selection systems are the

[8]http://www.trumanlibrary.org/oralhist/wright.htm

Table I: Summary of the different personation technologies

		Visual	Auditory	Animation
Physical	Artificial	Androids - Screens	Biomechanical - Loudspeaker	-
	Human	Twins - Prosthetic makeup	Impersonation	Impersonation
Digital	Record-based	Image-based (Tampering)	Unit-selection (Tampering)	Cloning
	Model-based	CG (Synthesis)	SSS (Synthesis)	Autonomous

most used type of synthetic speech and are employed in many real-life applications in our everyday lives.

Synthesis: Many technologies for speech synthesis rely on models of speech. These systems include but are not limited to: Statistical speech synthesis (SSS) [24], Articulatory speech synthesis, and voice conversion [25]. The most used type of synthetic speech generation systems is SSS. These systems model the distribution of speech features using HMMs in a similar manner to speech recognition systems, and later synthesize speech using parameter generation algorithm. The resulting speech lacks the naturalness of the unit-selection systems, but has more cohesion, is more flexible, and can model the high-level and dynamic features of the speech to some extent. The similarly is also high as the synthesis parameters are generated from the distribution of speech features extracted from genuine speech. The possibility of speaker adaptation on these systems makes them a good candidate for automated personation attempts.

Another type of synthetic speech that requires attention is voice conversion. Given an audio signal from a target speaker, the system can learn a mapping from feature space of the personator to that of the target speaker. This model can later be used to convert the voice of a personator to the target speaker's voice. As of now, these systems lack naturalness in their generated audios but may improve tremendously as the technology advances.

Wavenet [26] represents another interesting type of speech synthesis system that relies on waveform synthesis rather than feature synthesis. The clear naturalness and resemblance of human speech using waveform synthesis is promising and can pass human judgment.

3) Animation source: Digital speech synthesis systems usually get a text as an input for generating the output audio. The text may be accompanied by affective information as well. The exception to this is the voice conversion systems that act in a similar manner as the motion capture systems.

C. Combinations

Multimodal personations require combination of visual and auditory modalities. This is challenging, as humans rely on both visemes and phonemes to understand speech, and thus are extremely sensitive to small disaccords between modalities. The technology of choice for each modality can vary depending on the application of the system. Of course, these techniques can be combined on each modality as well, producing "*hybrid*" personations. This can be done to take advantage of their fusion to reduce the need for extra modeling, avoiding artifacts, or reducing computational costs. Obfuscation may also be employed concurrently to achieve the same goals.

Table II: An estimate of detection difficulty of personation attempts for humans, along with generation cost approximation. (E: Easy, M: Moderate, H: Hard) Naturalness and similarity are estimated for Visual (V), aniMation (M), and Auditory (A) aspects.

		Naturality			Similarity			Creation Cost	
		V	M	A	V	M	A	Model	Prod.
Physical	Artificial	H	E	E	H	E	E	High	Low
	Human	H	H	H	M	H	H	Mod	Low
Digital	Record-based	H	H	H	H	M	H	Low	Low
	Model-based	M	M	M	H	H	H	Mod	Low

III. Detection Techniques

Different detection technologies are being developed stemming from fields of digital video forensics, biometrics, and fake news detection. Presentation attack detection technologies address the detection of physical attempts while tampering detection and computer-generated detection technologies provide solutions for the detection of digital tampering and synthesis attempts respectively (Fig. 1). Despite considerable achievements, to date, no generalized method for automated detection of such content has been developed [27]–[29].

A different approach in development is to utilize contextual and style-based information as well as relying on external sources of knowledge for verification of veracity of a piece of information [5]. Hitherto, the task of personation detection remains mostly a manual endeavor.

IV. Discussion

In this study, we attempted to categorize all the existing applicable technologies for audiovisual personation. The list of personation technologies can be summarized in Table I. It can be seen that the visual, auditory, and animation factors of a given entity can each one be created by a human, by modifying an existing record, or by synthesis from a model, and done independently of one another. This classification simplifies the description of any personation technology as well as the formulation of weaknesses and strengths of these methods.

An estimation of the detection difficulty of personation attempts for viewers is given in Table II. Given the difficulty of detecting record-based models, it can be concluded that

major risks of existing technologies are presented by these personations. A lower level of risk arises from modeling of reality by humans and model-based systems.

V. Future work

In this study, different techniques that are usable for personation are listed and explained. Future work consists of studying the risk assessment of these attacks and applicable detection technologies. Creation of a dataset based on this classification and objective evaluation of the performance of different detectors would be the next step.

References

[1] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," National Bureau of Economic Research, Inc, NBER Working Papers 23089, 2017.

[2] B. Nyhan and J. Reifler, "When corrections fail: The persistence of political misperceptions," *Political Behavior*, vol. 32, no. 2, pp. 303–330, Jun 2010.

[3] Z. Jin, G. J. Mysore, S. Diverdi, J. Lu, and A. Finkelstein, "Voco: Text-based insertion and replacement in audio narration," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 96:1–96:13, Jul. 2017.

[4] J. Thies, M. Zollhfer, M. Stamminger, C. Theobalt, and M. Niener, "Face2face: Real-time face capture and reenactment of rgb videos," in *CVPR'16*, June 2016, pp. 2387–2395.

[5] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *SIGKDD Explor. Newsl.*, vol. 19, no. 1, pp. 22–36, Sep. 2017.

[6] M. E. Clifton-James, *I was Monty's Double*. Panther Books, 1958.

[7] M. Mori, "The uncanny valley," *Energy*, vol. 7, no. 4, pp. 33–35, 1970.

[8] S. Nishio, H. Ishiguro, and N. Hagita, "Geminoid: Teleoperated android of an existing person," in *Humanoid robots: new developments*. InTech, 2007.

[9] G. Lucas and R. McCallum, "Star wars: Episode I the phantom menace," 1999.

[10] E. Bumiller, "A new set of bush twins appear at annual correspondents dinner," *The New York Times*, p. 1, 2006.

[11] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *SIGGRAPH '02*. New York, NY, USA: ACM, 2002, pp. 388–398.

[12] J. Oreck, "The matrix reloaded: Unplugged," 2004.

[13] M. Seymour, C. Evans, and K. Libreri, "Meet mike: epic avatars," in *ACM SIGGRAPH 2017 VR Village*. ACM, 2017, p. 12.

[14] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: Learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 95:1–95:13, Jul. 2017.

[15] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *SIGGRAPH '97*. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1997, pp. 353–360.

[16] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *SIGGRAPH '99*. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1999, pp. 187–194.

[17] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar, "Acquiring the reflectance field of a human face," in *SIGGRAPH '00*. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 2000, pp. 145–156.

[18] S. Saito, L. Wei, L. Hu, K. Nagano, and H. Li, "Photorealistic facial texture inference using deep neural networks," *CoRR*, vol. abs/1612.00523, 2016.

[19] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: A back-projected human-like robot head for multiparty human-machine interaction," in *COST'11*. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 114–130.

[20] C. Cao, Q. Hou, and K. Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 43:1–43:10, Jul. 2014.

[21] F. I. Parke, "Computer generated animation of faces," in *Proceedings of the ACM Annual Conference - Volume 1*, ser. ACM '72. New York, NY, USA: ACM, 1972, pp. 451–457.

[22] K. Fukui, E. Shintaku, M. Honda, and A. Takanishi, "Mechanical vocal cord for anthropomorphic talking robot based on human biomechanical structure," *The Japan Society of Mechanical Engineers*, vol. 73, no. 734, pp. 112–118, 2007.

[23] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP'96*, vol. 1, May 1996, pp. 373–376 vol. 1.

[24] H. Zen, K. Tokuda, and A. W. Black, "Review: Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[25] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *ICASSP'88*, Apr 1988, pp. 655–658 vol.1.

[26] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[27] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Comput. Surv.*, vol. 50, no. 1, pp. 8:1–8:37, Mar. 2017.

[28] K. Sitara and B. Mehtre, "Digital video tampering detection: An overview of passive techniques," *Digital Investigation*, vol. 18, no. Supplement C, pp. 8 – 22, 2016.

[29] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, no. Supplement C, pp. 130 – 153, 2015.