Marc Böhlen

The Making of Fake Voices

Abstract: Im vorliegenden Text wird erörtert, wie technologische Innovationen und Fortschreibungen menschlicher Sehnsüchte die Sprachsynthese an einen Punkt geführt haben, an dem sie in industriellem Maßstab eingesetzt werden kann und dabei nahezu jede menschliche Stimme nachzubilden vermag. Der Beitrag trägt dem Unterschied zwischen perfekter Mimesis und erfahrungsbasierter maschineller Sprachproduktion Rechnung. Er legt dar, wie dieser Unterschied als Werkzeug der Täuschung eingesetzt werden kann, und wie er als ein Experimentierfeld dient, auf dem über das mittels künstlicher Intelligenz realisierbare Klonen menschlicher Eigenschaften im Allgemeinen nachgedacht wird.

Abstract: This text discusses how innovation in technology and continuity in human desires brought voice synthesis to a state in which it can be deployed at an industrial scale and reproduce almost any human voice. The text considers the difference between perfect mimesis and machinic speech production, describing how this difference can be deployed as a tool for deception, as well as the way it serves as a testing site for reflecting on artificial intelligence driven cloning of human features in general.

1 Introduction

In March 2019, criminals used artificial intelligence to impersonate a chief executive's voice for a fraudulent money transfer.[1] The scammers created a fake version of the voice of the chief executive and called an unlucky executive employee in the fake voice of this supervisor, which included a slight German accent and the specific melody of the supervisor's voice. The employee was informed by this fake boss that €200,000 were to be transferred to a compromised recipient address within an hour. The employee promptly executed the transaction, unwittingly enabling the first documented artificial intelligence generated fake voice cybercrime of the 21st century.

The history of synthetic speech spans at least three centuries (Ramsey 2019) and possibly much longer, if accounts of speaking statues as early as 20 BC informed

of principles described in Heron of Alexandria's treatises on machinery, mechanics and hydraulics (Pettorino 1999) are in fact true. Reflecting on the history of fake voices offers an opportunity to consider how one age-old dream can drive technical innovation across centuries. It can also serve as a case study in the downstream effects of technical innovation. While the phone scam example above might suggest that deception is a product of only the latest instantiation of artificial voice technology, con-artistry was in fact an early adapter to new communication opportunities afforded by landline telephony as it changed communication patterns and opened the door to new forms of impersonation (Marvin 1999).

By exploring several speech producing systems in context – Kempelen's Sprachmaschine, Dudley's Voder and Tacotron – this text will cast voice synthesis as a story of an immemorial human dream, implemented in each iteration utilizing the current technology available, and entangled with the social dynamics in which it is inserted. The last section then reflects on how we live with synthetic speech systems today under these entanglements.

2 Kempelen's Sprachmaschine

When Wolfgang von Kempelen began experimenting with a device that could imitate human voices, he already had some good reference points. By the beginning of the 17th century, a low fidelity mechanical model of how sound is generated in the human vocal tract had already been established (Ramsey 2019, p. 11). Kempelen's device translated the 17th century state of the art model of the human voice tract into a mechanical apparatus made of wood, paper, brass wires, tin tubes and leather. Yet the simplicity of the construction belies the depth of its potency. Kempelen produced a 464-page manuscript (Kempelen 1790) that not only offers detailed engineering drawings of the various mechanical parts of the apparatus and a lengthy treatise on the body parts involved in the production of speech, but also a general discussion on human language and its presumed origins.

As opposed to earlier attempts at voice-like sound creation, Kempelen's Sprachmaschine was the first device capable of generating utterances reminiscent of entire words.[2] Kempelen translated human sound production into an equivalent non-human system with a bellows functioning as lungs, a wind chest to distribute the air to sound producing enclosures, a reed made of a thin strip of ivory glued to a piece of leather, and a funnel made of natural rubber representing the oral cavity (Kempelen 1790, chapter 5; Deutsches Museum 2020). The device was more a musical instrument than a utilitarian apparatus. It was played by pressing on the

bellows and opening and closing pathways to enable or constrain airflow to the different parts of the machine. Under the skillful control of an operator (Braskhane 2017), a variety of human-like sounds could be produced to imitate short utterances in several different languages (Pettorino 2015) including Latin, French and Italian but not German and no report on Hungarian, Kempelen's mother tongue.

While Kempelen's well-known chess-playing automaton, the Mechanical Turk, was a clever mechanical contraption capable of moving chess pieces across a chess board, it was not a chess champion. In fact, Kempelen's Mechanical Turk was a fraud – there was a skilled human chess player inside the machine performing the chess moves out of view of the audience. No such post-mortem disclosure blemishes the Sprachmaschine. This is surprising given the fact that Kempelen presented both his Mechanical Turk as well as his Sprachmaschine together on tour across Europe in 1783 and 1784 (Deutsches Museum 2020). Moreover, experimentation in human voice creation up to the 17th century was generally viewed with suspicion and often accused of sorcery, persecuted and even condemned (Pettorino 2015). To give voice to an object was perceived as more amazing than to have it emit a melody; to give voice meant to give (humanlike) life to inanimate matter, a feat considered beyond the reach of human agency.

Despite and because of these circumstances, Kempelen's machine is a landmark in the history of voice generation. As a product of the Enlightment period, it mirrors a world view that perceived the human body as a machine. Kempelen's creation is the first viable construction capable of imitating the sound production of the human voice tract based on a mechanical model of this very voice tract. As a disembodied voice it represents one early example of a trajectory of scientific inquiry and engineering design that seeks to replicate human abilities and features operating outside of and without the need of the human body.

3 Dudley's Voder

When Homer Dudley conceived his speech apparatus in the middle of the 20th century, he also relied on the materials and techniques of his time. But instead of wood and leather, Dudley, a researcher at Bell Labs, assembled his device utilizing the hardware du jour, vacuum tube electronics.

In *The Carrier Nature of Speech* (Dudley 1940), Dudley outlines his speech generation concept as a carrier circuit, informed by the model of analogue radio communication. The carrier circuit describes an information representation concept in which a

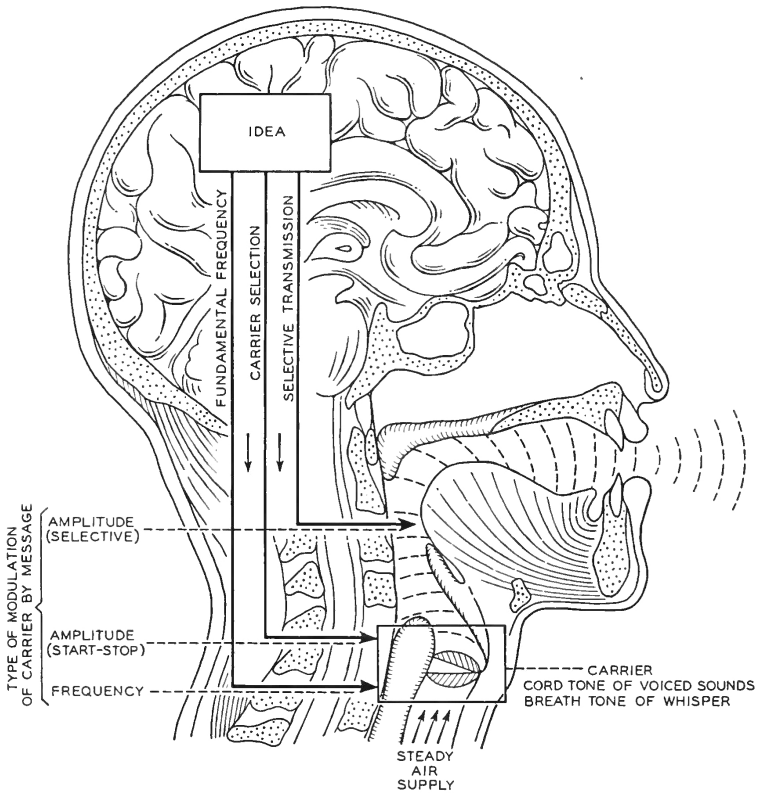

Fig. 1: From Homer Dudley's *The Carrier Nature of Speech*. The Bell System Technical Journal Vol 19, Number 4, October 1940, p. 497. Reused with permission of Nokia Corporation and AT&T Archives.

'transport' waveform is modified (modulated) with an information-dependent signal, usually higher in frequency than the base carrier wave. Dudley's concept maps speech produced by the dynamics of compressed air in the human vocal tract onto corresponding frequency bands. By selectively combining these frequency bands in the spectrum of human speech with a base carrier wave, Dudley was able to devise a voice synthesis approach capable of producing human-like speech. This result seemed rather counterintuitive as the carrier signal itself, sounding like a hiss or a

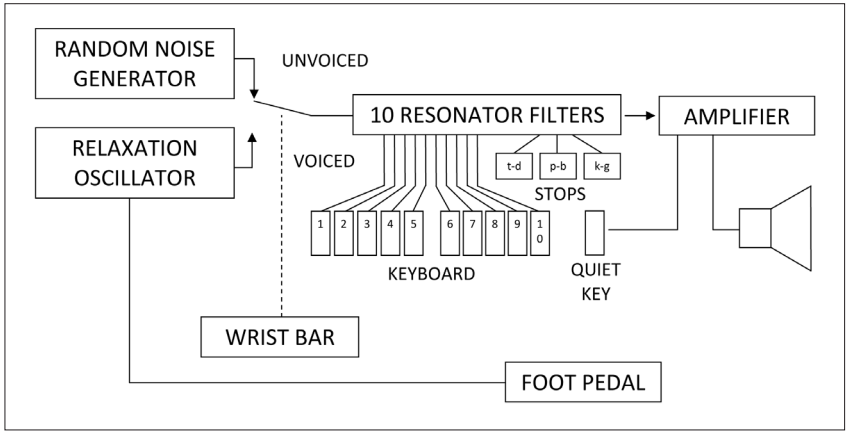

Fig. 2: Schematic diagram of the Voder.

buzz, is in no way reminiscent of a human voice; only the modulated product sounds anything like a human voice. A far cry from the kind of synthetic speech we have grown used to today, Dudley's approach was flexible enough to create both voiced utterances emulating sounds produced when the vocal cords vibrate (such as 'z') as well as unvoiced utterances (such as 's') in English.

Similar to the work of Kempelen, Dudley's concept included the source of sound creation. Dudley's concept integrates the human thought process into the model, including the idea that originates in the mind, and which only later is translated into the sound-producing hardware of the human body (see Fig. 1). Unlike Kempelen, Dudley's experiments limited themselves to the English language. However, Dudley at least reflected on other possible applications ranging "from the puffs of a locomotive to instrumental music" (Dudley 1940, p. 513).

This connection between idea and speech act continues to be pursued in current research under the theme of concept-to-speech recognition, formally defined as "the production of synthetic speech on the basis of pragmatic, semantic, and discourse knowledge" (Alter 1997, p. 4). Compared to Kempelen and Dudley's machinery, concept-to-speech is comparatively pedestrian in its ambitions, but it delivers tangible results, improving speech dialogue between computer-generated voices and people. For example, concept-to-speech has more recently been deployed on an industrial scale in Amazon's popular Alexa assistant.[3]

Dudley applied this carrier concept to an operator-controlled voice machine called the Voder (Voice Operation Demonstrator). The Voder produced a carrier wave with a buzzer-like sound for the voiced, and a hiss-like sound for unvoiced sounds. The Voder had a bank of 10 pre-defined band pass filters covering (most of) the spectrum of human speech. All these filters receive input from the noise source or the relaxation oscillator (the buzz source). The operator selects between these two input sources (carriers) with a wrist bar and controls the pitch of the input with a foot pedal. A keyboard acts as a controller on the filters, reducing or increasing the contribution of any one of them (see Fig. 2). Together with a quiet key, these components allowed an operator to play the device and generate sounds using different pitches and inflections that could be recognized as speech.

The Voder was not easy to utilize. Multiple operators had to train for over a year to be able to produce only a few simple utterances (Guernsey 2001). Just as Kempelen toured his machine to impress the crowds, Dudley's Voder was prominently featured at New York World Fair of 1939. Tellingly, that World Fair exhibited another main attraction capable of otherworldly speech, namely the robot Elektro (Marsh 2018). Elektro was a two-meter-tall humanoid robot that could walk upon spoken command, responding to the pattern of sounds from an operator – but not the content of the message. Moreover, Elektro could speak several hundred words prerecorded on a record player and differentiate between red and green colors with the help of a photoelectric camera eye. For these reasons, including the fact that Elektro would also smoke a cigarette, the robot's live performances captivated the World Fair's audience.

4 Speech Synthesis and Linguistics

Dudley's approach relied on the emulation of spectral patterns of human speech. The next steps in the history of speech synthesis required a more abstract approach – an approach that was simultaneously grounded in the theory of signal processing, guided by models of the vocal tract and Dudley's results, and informed by the insights of linguistics research, a field that hitherto was not a formal part of the synthetic speech research landscape.

Linguistics describes languages in generative terms with the goal of specifying rules for the generation of legitimate sentences through an abstract representation. Moreover, linguists represent spoken language using discrete elements, like language specific phonemes with particular features such as labial and nasal characteristics (Klatt 1987). Particular rules are then devised to explain when

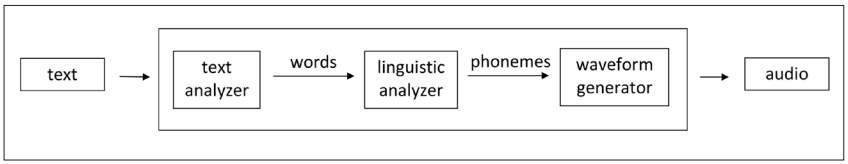

Fig. 3: Speech synthesis from text; diagram of the processing pipeline including text analysis, linguistic analysis and sound generation.

words change pronunciation in some sentence contexts (and not in others). These language specific rules – formalized in the text analyzer and linguistic analyzer – encode how a sequence of letters is transformed into a sequence of sound primitives that a waveform generator then assembles to an audible output (see Fig. 3). This rule-based approach emerged as an alternative to the more intuitive data-centric process of simply collecting a set of pre-recorded messages, and then combining elements of those pre-recorded segments to new utterances. Over time, two main approaches emerged to address the constraints and tradeoffs of the rule-centric vs the data-centric approach: formant synthesis and concatenative synthesis.[4]

5 Formant and Concatenative Synthesis

Formant synthesis is largely rule-driven. The synthesized speech is generated using an acoustic model and hand-crafted acoustic rules. Formant text-to-speech (TTS) creates speech segments from written text by generating signals based on language specific rules combined with general spectral properties of human speech. Formant TTS uses additive synthesis under the constraints of an acoustic model that describes the fundamental frequency, intonation, and prosody – the elements of speech that define individual articulation including tone of voice and accent.

Formant based methods can alter many aspects of a synthetic voice, including intonation, without relying on additional data. Because it is less dependent on data, formant TTS is ideal for gadgets, toys and household appliances where memory and processing power are limited. However, formant TTS is often recognizably machinic and is prone to glitches even when producing simple words; a condition often experienced when listening to directions uttered by formant TTS in early GPS navigation devices, for example.

Concatenative synthesis, instead, is data-driven. It relies on high fidelity audio recordings, from which segments are selected and combined via unit-selection

(selection of phonemes annotated with contextual information) to form a new speech utterance (Hunt 1996). Typically, a voice actor records several hours of speech which are then processed into a large speaker specific database containing linguistic units, phonemes, phrases and sentences. When speech synthesis is initiated, a speech generator searches this database for speech units that match those extracted from an input text, and concatenates these segments to produce an audible output. Concatenative TTS can produce high quality audio if a large and varied dataset has been collected. However, the approach makes it difficult to modify the voice (i.e. switching to a different speaker, or changing the emphasis or emotion of the speech) without recording a new database of phrases.

Both formant and concatenative TTS are in principle capable of constructing and uttering grammatically correct speech. However, both systems struggle with prosody, the subjective payload of speech. As such, neither approach is able to reproduce nuanced and sophisticated aspects of emphatic or emotional human speech across multiple languages and language use scenarios.

Despite these serious limitations, reports on lifelike synthetic speech periodically surface. As early as 1972 researchers reported on speech synthesis results which were so believable that listeners could not tell the difference between the synthetic and the human version, if presented in sequence (Klatt 1987, p. 743). One can assume that listeners in the 1970s might have been less discerning than they are today. Tellingly, the (male) researchers already then focused on the synthesis of male voices as they found the task of synthesizing a woman's (or a child's) voice more difficult (Klatt 1987, p. 746).

6 Speech Synthesis and Deep Learning

Rule-based and data-centric synthesis are not mutually exclusive, and machine learning in fact combines insights from both approaches. Deep learning is machine learning implemented with large scale, multi-layer (hence deep) network configurations, and deep learning TTS (or neural TTS) is the current preferred paradigm for designing synthetic speech systems.

Neural networks learn patterns from data. In speech synthesis, neural networks learn patterns from audio files. Once these patterns have been internalized in an iterative process, neural nets can create utterances that sound like the voices they have been exposed to. While neural TTS enables much more efficient adaptation

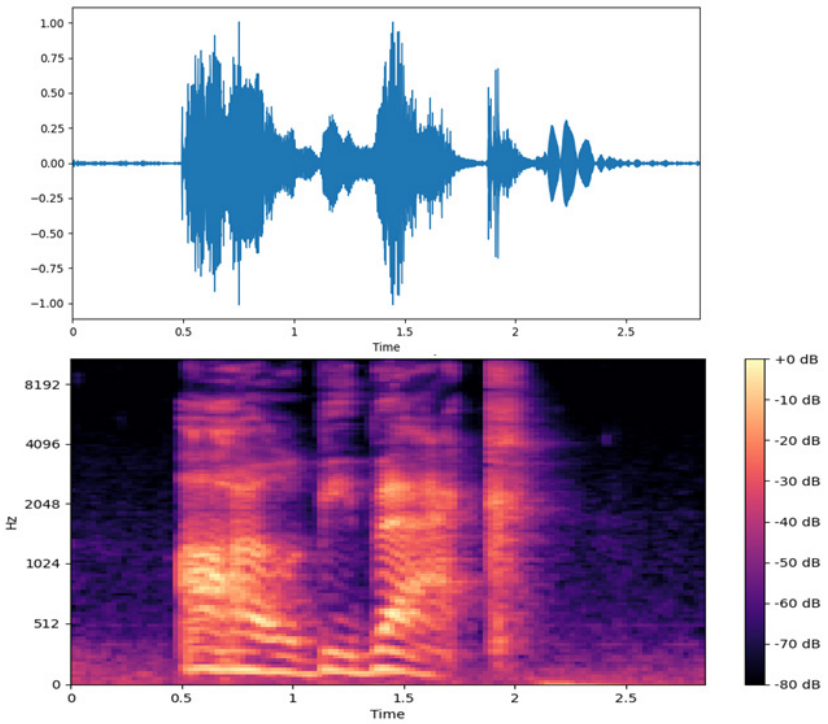

Fig. 4: Sound wave (top) and Mel spectrogram (bottom) of the three second utterance "I will be back" recorded by the author in English with a Swiss German accent. Left axis of the Mel spectrogram indicates frequency range normalized for human perception (pitch of equal distance on the scale 'sound' equally distant from each other). Right axis indicates intensity in decibels.

to new voices, variation in speaking patterns and expressive speech, neural TTS machinery comes – as engineering design does – with its own baggage.

Neural nets learn patterns from data in sequence in a process called training. Training refers to the iterative reducing of the distance (error) between output scores and the desired pattern of scores. The machine modifies its internal parameters (weights) across the various network layers to reduce this error, evaluates the outcome, then tries again, until the error is small. The learning algorithm computes

a function (gradient vector) for each weight indicating how the error would change if the weights were increased by a small value (LeCun 2015). Adjustment of the weights then occurs in the opposite direction of that gradient result, operating in an adaptive loop until the average value of the objective function stops decreasing. This adjustment is the magic sauce of the learning operation.

In neural network-based speech applications, input data appear as spectrograms created from the text. A spectrogram is a two-dimensional map of the frequencies that make up the sound, from low to high, as well as the changes of these frequencies over time, from left to right (see Fig. 4). Spectrograms are rich descriptors of text-voice constructions, supplying neural nets with detailed signal data to learn from while remaining oblivious to the messages contained in those signals.

Neural TTS systems allow for flexibility with fidelity unattainable through previous approaches. Changing the perceived gender of a voice, as well as building speech utterances that imitate a particular person with only a few examples of their speech patterns become routine operations. And this flexibility is precisely what the fake voice industry puts to nefarious use.

6.1 Tacotron

Most neural TTS systems combine two neural networks, one dedicated to translating text to a frequency representation, and a second one that converts that output to a synthetic voice. The basic principles of neural TTS are best described with an example.

Tacotron (the newest version at the time of this writing is Tacotron2) is a text to utterance generative model that synthesizes speech directly from characters. Given <text, audio> pairs, the model can be trained from scratch and delivers realistic results even to current discerning listeners. Tacotron is composed of two connected neural networks. The first is a recurrent feature prediction neural network that maps character embeddings to spectrograms. The second generates audible waveforms from those spectrograms (see Fig. 5). The first neural network has two main components, an encoder and a decoder. The encoder converts a character sequence into a feature representation which the decoder consumes to predict a spectrogram one frame at a time, capturing not only pronunciation of words, but also various subtleties of human speech, including volume, speed and intonation (Shen 2018). The second neural network is a WaveNet model, one that generates raw audio waveforms. It acts as the vocoder, the component that produces the actual

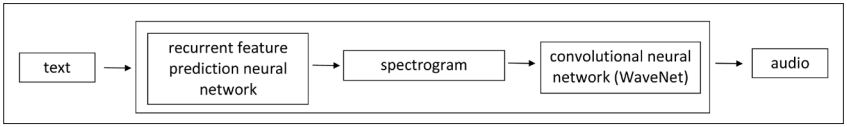

Fig. 5: Schematic diagram of the main components of Tactotron2 model with the WaveNet element. The prediction network performs the text and linguistic analysis while the convolutional network performs the work of waveform modelling.

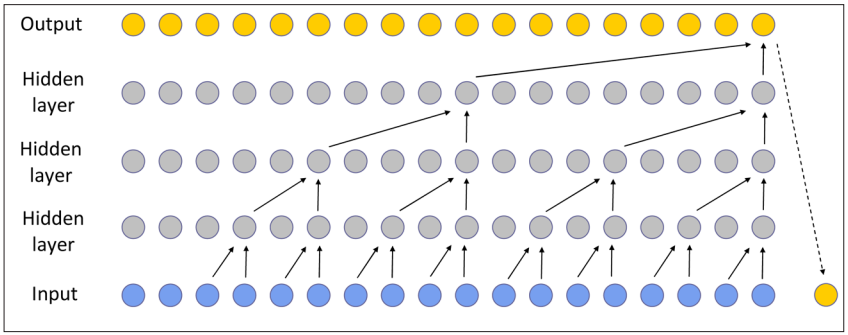

Fig. 6: Diagram of the casual convolution layers of WaveNet. After each sample is predicted as an output, it is fed back into the network as part of the input stream to predict the next sample (dashed arrow from top to bottom right most dot). This allows the receptive field to grow and cover continuous time-stepped inputs of a speech act. After Oord 2016.

sound not unlike the Dudley apparatus described above. However, this vocoder is trained iteratively on the spectrogram data produced from the first neural network. WaveNet ensures that the ordering of the time series audio data is preserved in the learning process. The model's predictions unfold sequentially: after each sample is predicted, it is fed back into the network (Oord 2016) to predict the next sample (see Fig. 6). This approach provides a high degree of flexibility. If one trains WaveNet on American English, it produces American English speech, if you train it on German, it produces German. As such, WaveNet is a universal speech engine; it models spoken language through its neural logic dynamics, absorbing whatever sound patterns it is subject to during training. However, it cannot discern whether parts of the sound landscape are relevant or not, leading to instances in which background sounds from the room in which recordings occurred were imitated (House 2017). As a neural learning machine, WaveNet is completely dependent on its training

data yet it creates from this assemblage a new form of acoustic knowledge (House 2017, p. 20). Once trained, WaveNet's knowledge is stored in the parameters of its model, which can be tuned to control the characteristics of a speech act, making the WaveNet architecture an ideal candidate for voice cloning as described below.

Neural TTS has markedly reduced the difference between machine-generated and human produced speech, and most major global IT companies including Nvidia (WaveGlow), Deepmind (WaveNet), Mozilla (LPCNet) and Baidu (DeepVoice) have developed proprietary neural TTS systems. Voice-based interaction is big business.

7 Living with Fake Voices

Early TTS, both of the formant and concatenative variety, have been deployed in a plethora of gadgets, appliances, and navigation aids, creating a vibrant ecology of first-generation fake voices. There was little opposition to the expanding collection of these early computer voices as their reach was limited. In fact, system imperfections made them quirky yet recognizable as non-human. As such they allowed human beings to navigate TTS-infused interaction events with call center agents and robot bank tellers. Deviations from the baseline of human naturalness served as a form of auditory landmarks that formed demarcation points between service robots and humans. Those not-quite-real synthetic voices were well-defined as non-human, non-threatening additions to a world ruled by human beings. Neural TTS changes this condition because of almost undetectable deviations from human speech, and because of the scale at which neural TTS is deployed; from personal mobile phones to platform products, all computer systems now support voice interaction. Neural TTS can even emulate bodily speech features such as lip smacking,[5] reintroducing a window onto materiality previously obscured, and suggesting believably that a living, breathing body is in fact producing its speech acts. As such, neural TTS destabilizes established frameworks that allow humans to identify computers in action and introduces new flavors of uncanniness into computer interaction.

7.1 Depressed Voice Talent

It is maybe not without irony that state-of-the-neural-art voice synthesis systems require copious amounts of data to achieve their superior performance. And this data comes invariably from real people, acting as voice talent in the parlance of the speech industry. The term voice talent heralds from the performing arts and radio production in which the significance of a well-balanced and articulate voice

was long appreciated. Clarity of pronunciation and clear articulation are equally significant for voice synthesis, and so performing arts voice talents delivered the first set of voices to the fake voice industry. Siri's US voice is based on the voice of voice performer Susan Bennett, and the voice of Cortana is based on Jen Taylor's.[6] Siri was originally conceived as an 'assistant' but has left those humble origins behind and has been integrated into popular culture through appearances in television shows. The voice itself has become a household name and a pop star, despite having no physical connection to the human being who sourced its famous sound. So pervasive is the pull of Siri as a cultural phenomenon that the originating human, Mrs. Bennett, became in turn accidentally famous and a popular speaker reporting on "what it is like to be the person behind Siri"[7]. The relationship between synthetic voices and their human sources is a fragile one, with some voice actors reporting a sense of disappointment and sadness after being replaced and updated by a more fashionable voice (cf. note 6).

While synthetic voices now sound largely realistic, the social realities they represent have remained stubbornly conservative. The Matthews, Johns, Kendras and Sandras of the speech industry suggest an identity binarized world, and the voice flavors remain strictly either male or female. Moreover, the majority of voices designed for assistant tasks are female or female by default (UNESCO 2019). Before this biased voice landscape was recognized as problematic by industry, artists, including the author of this text, investigated pathways by which to probe the language normalization in synthetic voice design, including the construction of immigrant accented language (Böhlen 2008).

Only recently, the synthetic voice industry has responded to gender fluidity in the construction of an appropriate synthetic voice. Sam is the world's first comprehensive non-binary voice product.[8] Sam differs from previous gender-neutral voice products such as Q[9] that attempt to avoid gender specific characteristics altogether and appear genderless, in that it combines prosody features from both male and female voices to a voice product that sounds like a man *and* a woman. Sam is marketed specifically to industry products seeking to resonate with the transgender or gender non-conforming community (cf. note 8). In fact, the newfound flexibility in voice product fine-tuning allows to design voices sounding in any manner one might desire; with the technological hurdles removed, voice developers explore the edges of voice design and have recently arrived at two strange places, to wit voice cloning and deep fakes.

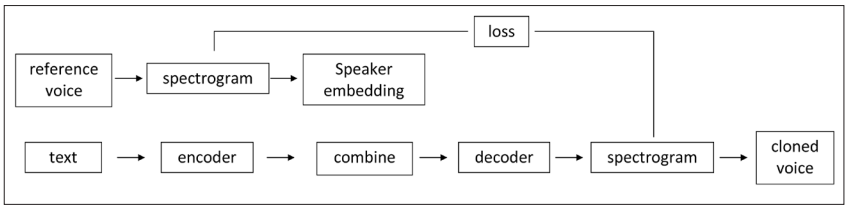

Fig. 7: An individual's unique voice characteristics are added to a neural TTS system via speaker embedding. After Jia 2019.

7.2 Voice Cloning

Voice cloning creates a synthetic voice of a specific (living or dead) human being. The unique tonal characteristics of a living person's voice can be captured through recording samples and transferred as speaker embeddings (see Fig. 7) into a neural TTS system. As such, voice cloning relies on a listening operation before it becomes a speaking machine.

Voice cloning extracts the salient features of a speaker's voice from a reference audio utterance in order to create the speaker embedding, paying no attention to the meaning in the utterances and collecting only characteristics such as pitch, accent and tone.

The speaker embedding information is combined with the phoneme sequence, and the vocoder generates from this combination a voice with the auditory features of the specific speaker, in other words the cloned voice. The neural magic learning sauce that fine-tunes how closely the voice sounds to the original recording occurs by comparing the (spectrogram of the) original recording with the spectrogram created by the decoder, and iterating through the process, making small adjustments until the loss, the difference between the reference and the clone, is negligible as described in the section on neural network training above.

Voice cloning finds a direct application in the form of voice banking, the collection of audio recordings for voice synthesis in anticipation of a voiceless future due to the consequences of degenerative illnesses such as *amyotrophic lateral sclerosis,*[10] or in order to counter adverse effects of surgical inventions such as laryngectomy. The entertainment industry has its own interest in voice cloning. Film narration for global audiences, podcasting, as well as game character narration all make use of voice cloning. Given the proliferation of software tools that facilitate the programming

of neural TTS and the opportunity to replace studio with home computer audio recordings,[11] the bar for creating cloned voices is at the present comparatively low, unleashing a wave of low-cost, cross-language product branding scenarios.[12]

7.3 Audio Deep Fakes

Audio deep fakes make use of the voice cloning techniques outlined above. However, audio deep fakes are created with sound recordings collected without consent. Typically, voice samples are collected from recorded speeches, public presentations, interviews or press conferences and used to train a voice cloning system. While voice cloning requires copious amounts of data for highest-quality reproductions and longer utterances, voice cloning systems applied to simpler tasks and short statements can operate with as little as 10 seconds of reference audio recording (Chen 2019) and still produce believable results. Even a phone call, recorded in a relatively noisy location, can be used as the source of voice-embedding and can clone a person's audio footprint, albeit with low fidelity. Scammers cleverly respond to these limitations by deploying audio deep fakes in applications in which one would in fact expect to hear low-quality audio, such as in telephone calls. The noisier the environment, the more difficult it is to distinguish a fake from a real voice. Because we are accustomed to hearing low-quality audio on telephones, low-quality audio fakes encounter a reduced threshold for scrutiny. It is precisely this combination of ubiquitous voice cloning software, busy conference calling culture and currently deficient literacy in recognizing voice fakery that enabled the first audio deep fake cybercrime described at the beginning of this text.

7.4 Audio Deep Fake Fraud

Electronically enabled fraud is at least as old as the landline telephone. Con artists used the public's unfamiliarity with voices heard over low-bandwidth telephone lines to impersonate other people or to gain trust in ways that would never have been possible in a face-to-face encounter (Marvin 1990). Likewise, "419 email" frauds in which an unsuspecting victim receives an email offering an "opportunity to share in a percentage of a large sum of money"[13] have been effective only because the trust landscape of email, particularly for the non-digitally native, is still unstable, functioning simultaneously as a personal communication channel as well as an official source of information.

Audio deep fakes likewise make use of the current unstable status of ubiquitous synthetic voice systems. Robotic-sounding voices are recognized as such and even enjoyed for their style; non-consensual individual-imitating voices do not yet have a defined space in everyday life. We have insufficient training to detect the new minor slippages that give them away, and when we are confronted with these new audio artifacts embedded in mundane situations, there is often no reason to be suspicious.

Since there is no public literacy in synthetic voice misuse so far, substantial efforts are being deployed to recognize deep fake audio through technical means. Assessing the veracity of a deep fake audio artifact requires authentication, i.e. proving it is fake or showing that it is not fake. Voice biometrics (Gonzalez 2008) offers heuristics by which to assess whether a voice stems from a genuine flesh and blood human being or from an algorithm. Anomaly detection, for example, can reveal if sounds in an audio artifact were generated through the anatomy of the human vocal tract or not (Hill-Wilson 2020).

Legal scholarship is also responding to the practice of speech cloning, as it attempts to discern whether synthetic speech is a form of protected speech. Likewise, the issue of copyright within voice cloning remains unresolved. Deep fakes use a variety of audio recordings, and these are processed to exhibit features that are no longer reminiscent of the original speech act, hence there is no direct copying in this approach, but rather an artistic re-use more in tune with the established concept of fair practice and derivative work for parody, for example in the Bern Convention and the US Copyright Law; a distinction the entrepreneurial producer Jay-Z seems to dismiss in his lawsuit against the YouTube channel *Vocal Synthesis*[14] that pokes fun at many celebrity figures, including Jay-Z, with neural TTS produced fake voice-overs.[15]

Legal scholars are particularly concerned with deep fakes at the far end of the sophistication spectrum as they undermine the reliability of genuine evidence (Pfefferkorn 2020). Moreover, they offer fraudsters a new avenue for deception and mistrust as *any* audio recording can now be declared as potentially fake and hence any evidence declared as invalid, simply because these new methods of fakery exist in general. The space of potential frauds, particularly when expanded by interaction with voice assistants, appears substantial. One can imagine a con artist manipulating a voice assistant with bogus claims (Lang 2018) uttered in a pain-ridden voice: “Alexa I have a terrible headache; please order some Aspirin”, to plant a history of fake evidence for a future insurance claim that could eventually be audited.

7.5 An Uncanny Valley of Fake Voices

The uncanny valley is a coinage used by roboticists describing "the proposed relation between the human likeness of an entity and the perceiver's affinity for it" (Mori 1970), or more practically, the experience of encountering a humanoid robot, only to notice that the robot is in fact not a person, but rather a 'fake' acting like a real person. When applied to the realm of synthetic voices, uncanniness is a product both of how the robot voice sounds as well as how the interaction dialog between the robot and the human unfolds. Research on chatbot artificial intelligence suggests that people in fact experience lesser uncanny effects and less negative affect when cooperating with simple text chatbots as opposed to more elaborate visually animated chatbots (Ciechanowski 2019). Moreover, the finer details of prosody of speech become all the more important in extending believability and preventing negative affect typical of the uncanny valley experience. Failing to differentiate tonality between short and long sentences – a task humans handle with grace – is hard even for neural TTS systems. Such slippage impacts the context of what is being expressed, and that mismatch becomes a source of negative affect (Simon 2019). The closer synthetic speech approaches real human generated speech while failing to perfectly replicate it, the stronger the sense of discomfort.

Perhaps the recent progress in near perfect voice imitation has blinded researchers to approaches that navigate the uncanny valley with more aplomb. The problem has been informally addressed in popular science fiction through a variety of robot characters. R2-D2 from *Star Wars* (USA 1977, D: George Lucas) delineates the human-robot boundary by speaking in a tongue no human can understand. The emotional connection between the robot and human beings is established instead through movements and gestures. Moreover, when R2-D2 becomes agitated, it produces wild steams of electronic sounds that unambiguously convey its inner state without ever saying a word. HAL 9000, the infamous disembodied artificial intelligence agent in Stanley Kubrick's film *2001: A Space Odyssey* (UK/USA 1968), on the other hand, tries tirelessly to sound precisely like a human being. And as its true intentions become apparent over the course of the film, HAL's more human-like voice renders its machinations all the more sinister.

Perhaps the best example of the handling of the inevitable mismatch between imitation and real human voice is offered by Stephen Hawking. Even as advanced voice synthesizers became available, the late physicist famously preferred an antiquated formant synthesizer.[16] "Perfect Paul", as the voice was coined, was a thin, tinny voice that became Stephen Hawking's widely recognizable vocal style.

Hawking maintained this technologically outdated voice even as his other assistive technologies were updated in order to accommodate his particular needs.[17]

7.6 When New Technologies Are Old (Again)

Prior to ubiquitous landline telephony, the concept of presence required the physical colocation of people (Marvin 1990). The telephone changed that logic of presence and replaced it with one of temporal continuity. Yet the change took some getting used to. It was only over time that the physical co-location requirement was sufficiently relaxed in a way that the general public understood that a person connected via telephone line could be addressed as if they were in the same room. Synthetic voices inherit the paradigm of presence from afar. They add to this condition a new twist, navigating (from afar) the presence of entities that sound like real people.

The uptake and governance of technological change is subject to many factors. The Collingridge dilemma describes the quandary according to which control of technology is difficult at early stages as not enough is understood about its consequences, and costly later on once the consequences are in fact apparent (Collingridge 1980). The development of synthetic voices offers the opportunity to add more nuance to the dilemma. Synthetic speech has long existed as an *old* new technology, one that prepared us for its arrival in imaginaries and crude approximations, appearing several times, in different forms, and eventually becoming globally distributed in systems small and large.

Synthetic voices no longer operate as isolated artifacts. Having positioned themselves into close contact with human beings in daily life, they have become dependent on and subject to changing dynamics of human social interaction and periodically adaptable to those dynamics, as the development of non-binary voice products demonstrates.

Synthetic voice technology is the first technology that can make a claim to have recreated human features so credibly that even human beings cannot distinguish the real from the artificial. More recently, portrait imagery has established a similar status with lifelike images of people who do not exist but were instead created by neural networks (Karras 2019). In both cases the neural techniques destabilize historically established concepts of veracity.

Observing how we now live alongside state-of-the-art voice systems serves as a testing ground for cohabiting with artifacts and systems based on human feature cloning in general. As artificial intelligence artifacts become more engrained in

everyday life, replacing older technical systems we have grown accustomed to – and introducing new unfamiliar ones – we will be tasked with developing new cohabitation strategies that allow us to monitor the effects of cloning-capable artificial intelligence on human wellbeing. New literacies and hitherto unfamiliar forms of connoisseurship of the artificial will likely emerge over time.

Locally contextualized and endlessly adaptable, pleasant sounding synthetic voices will make voice artists and announcement personnel redundant and positions at fast food restaurants obsolete. The always cheery voice agents that can handle routine customer requests with never ending artificial grace exhibit otherworldly patience, even in the face of the rudest of human beings. As such synthetic voices operate as early harbingers of technologies that replace human soft skills, previously immune to technology-driven labor displacement. For this reason alone, the evolving landscape of industrial synthetic voice deployment deserves attention, and careful listening. After all, none of these machines that sound like a human has any idea what it is talking about.

References

Alter, Kai; Pirker, Hannes; Finkler, Hannes (eds.) (1997): Introduction to the Workshop. In: Proceedings of a Workshop in Conjunction with 35th Annual Meeting of the Association for Computational Linguistics. https://www.aclweb.org/anthology/W97-1200.pdf [last accessed July 28, 2021].

Bern Convention (2020): Berne Convention for the Protection of Literary and Artistic Works. Article 10. https://treaties.un.org/doc/Publication/UNTS/Volume%20828/volume-828-I-11850-English.pdf [last accessed July 28, 2021]. https://www.wipo.int/export/sites/www/treaties/en/documents/pdf/berne.pdf [last accessed July 28, 2021].

Böhlen, Marc (2008): Robots with Bad Accents: Living with Synthetic Speech. In: Leonardo. 41/3. Pp. 209-214.

Boilard, Johnathan; Gournay, Philippe; Lefebvre, Roch (2019): A Literature Review of WaveNet: Theory, Application, and Optimization. In: 146th Convention of the Audio Engineering Society. Paper No. 10171.

Braskhane, Fabian (2017): The Speaking Machine of Wolfgang von Kempelen. https://www.youtube.com/watch?v=k_YUB_S6Gpo&ab_channel=FabianBrackhane [last accessed September 15, 2021].

Chen, Yutian; Assael, Yannis; Shillingford, Brendan; Budden, David; Reed, Scott; Zen, Heiga; Wang, Quan; Cobo, Luis; Trask, Andrew; Laurie, Ben; Gulcehre, Caglar; van den Oord, Aäron; Vinyals, Oriol; de Freitas, Nando (2019): Sample Efficient Adaptive Text-to-Speech. In: ICLR 2019. International Conference on Learning Representations. New Orleans, LA.

Ciechanowski, Leon; Przegalinska, Aleksandra; Magnuski, Mikolaj; Gloor, Peter (2019): In the Shades of the Uncanny Valley: An Experimental Study of Human-Chatbot Interaction. In: Future Generation Computer Systems. 92. Pp. 539-548.

Deutsches Museum (n.d.): Der Kempelen'sche Sprechapparat. https://www.deutsches-museum.de/forschung/forschungsbereiche/wissenschaftsgesch/sonic-visual-exhibit/sprechapparat/ [last accessed July 28, 2021].

Dudley, Homer (1940): The Carrier Nature of Speech. In: The Bell System Technical Journal. 19/4. Pp.495-516. https://archive.org/details/bellsystemtechni19amerrich [last accessed July 28, 2021].

González-Rodríguez, Joaquín; Toledano, Doroteo T.; Ortega-García, Javier (2008): Voice Biometrics. In: Handbook of Biometrics. Anil K. Jain; Patrick Flynn; Arun A. Ross (eds.). Boston, MA: Springer. Pp. 151-170.

Guernsey, Lisa (2001): The Desktop That Does Elvis. In: The New York Times. August 9. https://www.nytimes.com/2001/08/09/technology/the-desktop-that-does-elvis.html [last accessed July 28, 2021].

Hill-Wilson, Martin (2020): Keeping Contact Centers Secure as Fraud Intensifies and Gets Smarter. Investing in the Right Generation of Voice Biometrics. In: PinDrop whitepaper. https://www.pindrop.com/lp/white-papers/fraudster-journey-sep19/ [last accessed July 28, 2021].

House, Brian (2017): Machine Listening: Wavenet, Media Materialism and Rhythm Analysis. In: APRJA. 6/1. Pp. 16-24.

Hunt, Andrew; Black, Alan (1996): Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In: IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. Atlanta, GA, USA. Pp. 373-376.

Jia, Ye; Zhang, Yu; Weiss, Ron; Wang, Quan; Shen, Jonathan; Ren, Fei; Chen, Zhifeng; Nguyen, Patrick; Pang, Ruoming; Moreno, Ignacio L.; Wu, Yonghui (2018): Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis. In: 32nd Conference on Neural Information Processing Systems (NeurIPS 2018). Montréal, Canada.

Karras, Tero; Laine, Samuli; Aila, Timo (2019): A Style-Based Generator Architecture for Generative Adversarial Networks. In: CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA. Pp. 4396-4405.

von Kempelen, Wolfgang (1790 / 2017): Der Mechanismus der menschlichen Sprache. / The Mechanism of Human Speech. Fabian Brackhane; Richard Sproat; Jürgen Trouvain (eds.). Kommentierte Transliteration & Übertragung ins Englische / Commented Transliteration & Translation into English. Dresden: TUDpress.

Klatt, Dennis (1987): Review of Text-to-Speech Conversion for English. In: Journal of the Acoustical Society of America. 82/3. Pp. 737-793. https://acousticstoday.org/klatts-speech-synthesis-d/ [last accessed July 28, 2021].

Lang, Robert; Benessere, Lenore (2018): Alexa, Siri, Bixby, Google's Assistant, and Cortana Testifying in Court. Novel Use of Emerging Technology in Litigation. In: The Computer & Internet Lawyer. 35/7. Pp. 16-20.

LeCun, Yann; Bengio, Yoshua; Hinton, Geoffrey (2015): Deep Learning. In: Nature. 521/7553. Pp. 436-444.

Marsh, Allision (2018): Elektro the Moto-Man Had the Biggest Brain at the 1939 World's Fair. In: IEEE Spectrum. September 28. https://spectrum.ieee.org/tech-history/dawn-of-electronics/elektro-the-motoman-had-the-biggest-brain-at-the-1939-worlds-fair [last accessed July 28, 2021].

Marvin, Carolyn (1990): When Old Technologies Were New: Thinking About Electric Communication in the Late-Nineteenth Century. Oxford: Oxford University Press.

Mori, Masahiro (1970): The Uncanny Valley. In: Energy. 7/4. Pp. 33-35. English Translation. https://spectrum.ieee.org/automaton/robotics/humanoids/the-uncanny-valley [last accessed July 28, 2021].

van den Oord, Aaron; Dieleman, Sander; Zen, Heiga; Simonyan, Karen; Vinyals, Oriol; Graves, Alex; Kalchbrenner, Nal; Senior, Andrew; Kavukcuoglu, Koray (2016): WaveNet: A Generative Model for Raw Audio. In: 9th ISCA Speech Synthesis Workshop. September 13-15. Sunnyvale, USA.

Pettorino, Massimo (1999): Memnon the Vocal Statue. In: 14th International Congress of Phonetic Sciences (ICPhS-14). San Francisco, USA. Pp. 1321-1324.

Pettorino, Massimo (2015): The History of Talking Heads: The Trick and the Research. In: HSCR 2015 – Proceedings of the First International Workshop on the History of Speech Communication Research. Rüdiger Hoffmann; Jürgen Trouvain (eds.). Dresden: TUDpress. Pp. 30-41.

Pfefferkorn, Riana (2020): Deep Fakes in the Courtroom. In: Boston University Public Interest Law Journal. 29/2. Pp. 245-276.

Ping, Wei; Peng, Kainan; Gibiansky, Andrew; Arık, Sercan; Kannan, Ajay; Narang, Sharan (2018): Deep Voice 3: Scaling Text-to-Speech With Convolutional Sequence Learning. In: ICLR | 2018. Sixth International Conference on Learning Representations.

Ramsay, Gordon (2019): Mechanical Speech Synthesis in Early Talking Automata. In: Acoustical Society of America. Acoustics Today. 15/2. Pp. 11-19.

Shen, Jonathan; Pang, Ruoming; Weiss, Ron; Schuster, Mike; Jaitly, Navdeep; Yang, Zongheng; Chen, Zhifeng; Zhang, Yu; Wang, Yuxuan; Skerry-Ryan, RJ; Saurous, Rif; Agiomyrgiannakis, Yannis; Wu, Yonghui (2018): Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* April 15-20. Calgary, AB, Canada. Pp. 4779-4783.

Simon, Matt (2019): The Uncanny Valley Nobody's Talking About: Eerie Robot Voices. WIRED.com. March 18. https://www.wired.com/story/uncanny-valley-robot-voices/ [last accessed July 29, 2021].

US Copyright Law (n.d.): 17 U.S. Code § 103.Subject Matter of Copyright: Compilations and Derivative Works. https://www.govinfo.gov/content/pkg/STATUTE-90/pdf/STATUTE-90-Pg2541.pdf#page=5 [last accessed July 29, 2021].

West, Mark; Kraut, Rebecca; Chew, Han E. (2019): I'd Blush If I Could: Closing Gender Divides in Digital Skills Through Education. Think Piece 2: The Rise of Gendered AI and Its Troubling Repercussions. EQUALS and UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=85 [last accessed July 29, 2021].

Notes

1 https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 [last accessed July 29, 2021].

2 An original machine as well as a reconstruction are on view at the Deutsches Museum.

3 https://www.amazon.science/blog/new-text-to-speech-generator-and-rephraser-move-alexa-toward-concept-to-speech [last accessed July 29, 2021].

4 The story of synthetic speech synthesis in the post WW2 period into the 1990s is one of experimentation, failures and dead ends included. The interested reader may want to consult Klatt's 1987 *Text-to-Speech Conversion for English* for detailed accounts of these experiments.

5 https://www.scientificamerican.com/article/new-ai-tech-can-mimic-any-voice/ [last accessed July 29, 2021].

[6] https://www.theguardian.com/technology/2015/aug/12/siri-real-voices-apple-ios-assistant-jon-briggs-susan-bennett-karen-jacobsen [last accessed July 29, 2021].

[7] https://www.gossipcop.com/the-voice-of-siri/2560582 [last accessed July 29, 2021].

[8] https://newsroom.accenture.com/news/accenture-and-cereproc-introduce-and-open-source-the-worlds-first-comprehensive-non-binary-voice-solution.htm [last accessed July 29, 2021].

[9] https://www.genderlessvoice.com/about [last accessed July 29, 2021].

[10] https://www.als.org/navigating-als/resources/fyi-guide-voice-banking-services [last accessed July 29, 2021].

[11] https://www.cereproc.com/en/products/cerevoiceme [last accessed July 29, 2021].

[12] https://www.resemble.ai/ [last accessed July 29, 2021].

[13] https://www.fbi.gov/scams-and-safety/common-scams-and-crimes/nigerian-letter-or-419-fraud [last accessed July 29, 2021].

[14] https://www.youtube.com/channel/UCRt-fquxnij9wDnFJnpPS2Q/videos [last accessed July 29, 2021].

[15] https://www.theverge.com/2020/4/28/21240488/jay-z-deepfakes-roc-nation-youtube-removed-ai-copyright-impersonation [last accessed July 29, 2021].

[16] https://en.wikipedia.org/wiki/Talk%3ADECtalk [last accessed July 29, 2021]. That computationally elegant synthetic voice could operate within the constraints of an old minicomputer (an Intel 8086 chip).

[17] https://www.wired.com/2015/01/intel-gave-stephen-hawking-voice/ [last accessed July 29, 2021].

Der vorliegende Aufsatz entstammt der Publikation

Marcus Erbe / Aycha Riffi / Wolfgang Zielinski (Hrsg.)
Mediale Stimmentwürfe
Perspectives of Media Voice Designs

Schriftenreihe Digitale Gesellschaft NRW, Bd. 7
Kopaed Verlag, 2022

ISBN 978-3-96848-642-0