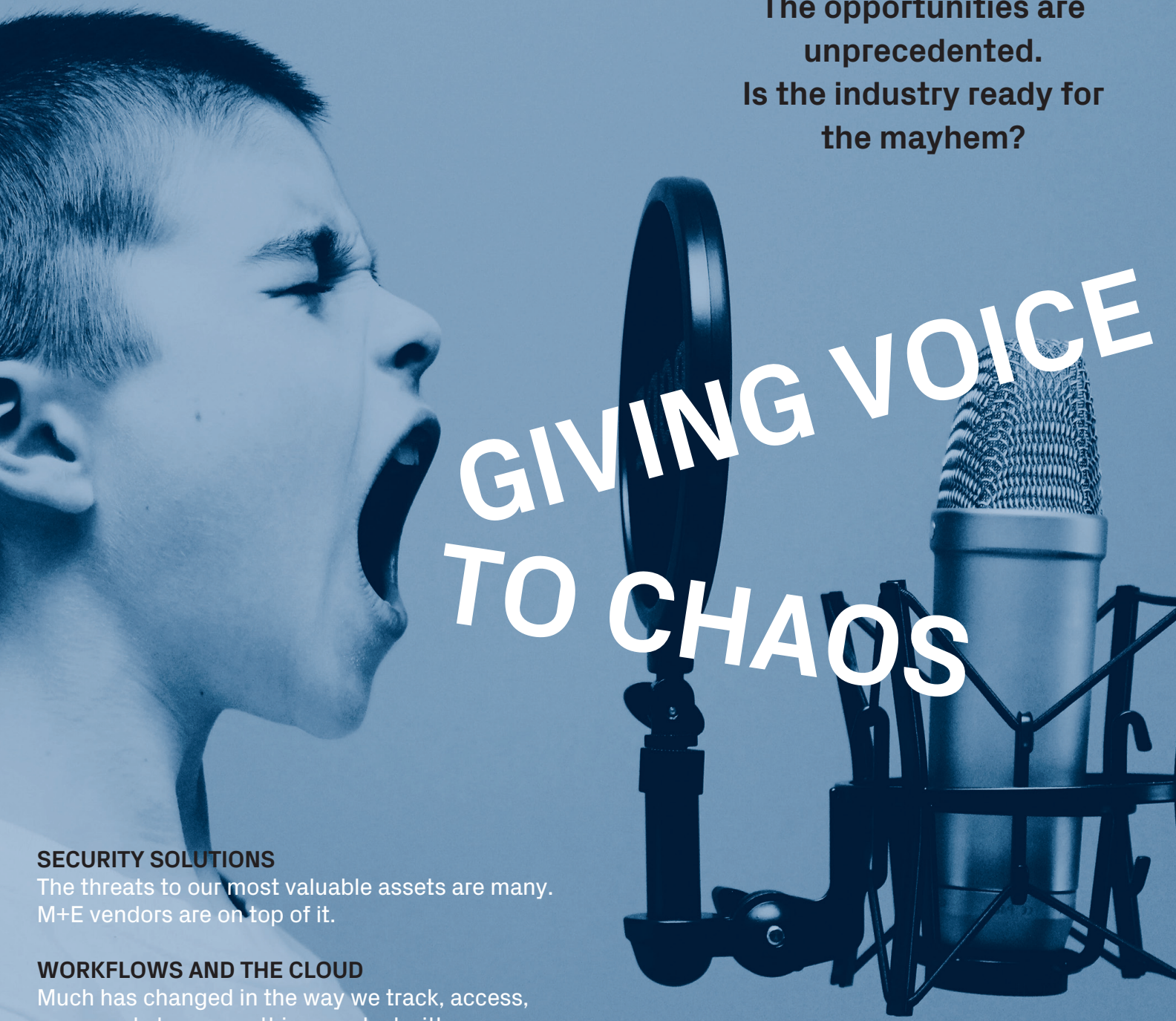
M+E JOURNAL

Today's localization challenges are enormous. The opportunities are unprecedented. Is the industry ready for the mayhem?



GIVING VOICE TO CHAOS

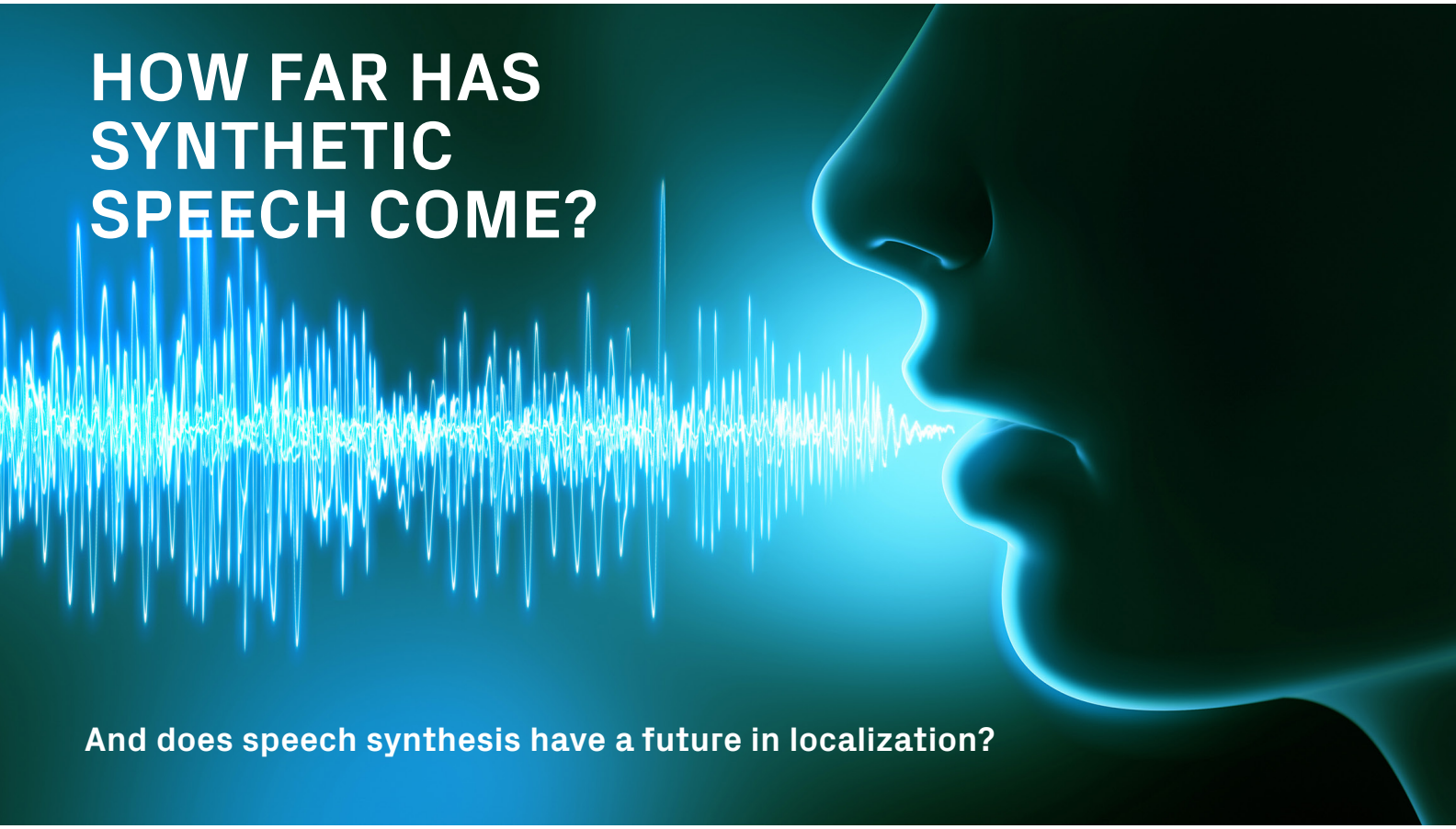
SECURITY SOLUTIONS
The threats to our most valuable assets are many. M+E vendors are on top of it.

WORKFLOWS AND THE CLOUD
Much has changed in the way we track, access, move and store everything we deal with.

SMART CONTENT
The many ways the industry adopts new technologies to make content smarter.

22.02

HOW FAR HAS SYNTHETIC SPEECH COME?



And does speech synthesis have a future in localization?

ABSTRACT: Speech synthesis, a technology that started as assistive to the blind, has entered our lives for good. As deep learning-based synthetic voices approach the naturalness of the human voice, new opportunities for the implementation of the technology are appearing. Digital assistants, web content, games and social media are now common ground for speech synthesis, but where else might we see it in the future?

By Volker Steinbiss, Managing Director, AppTek GmbH

Speech synthesis is a technology that has been in the forefront of many discussions recently in the media and entertainment industry, with a strong media presence and funding spree. It is already a large part of our lives in the form of Siri, Alexa, and other digital assistants, found in call centers, embedded in modern automobiles, or as assistive technology for the blind.

It is interesting then to understand what milestones the technology has undergone to reach the quality and the use cases for synthetic speech we are enjoying today, as well as think about where we might expect it to go in the future.

Originally synthetic voices did not sound natural as people tried to reproduce speech mechanically without much success. Computer-based approaches in text-to-speech began in the late 1960s and we can hear early examples of systems developed in the 1970s and 1980s here. Such systems were based on expert knowledge and explicit modeling of the vocal tract, and it was not until the 1990s that the first data-driven approaches were used: the concatenative approach and the statistical parametric speech synthesis, which utilized some machine learning in combination with Hidden Markov Models (also used in speech recognition at the time). Those voices were much easier to understand but still sounded somewhat robotic and were the norm for a long time, to such an extent that a large part of the public still associates synthetic speech with them.

Things changed drastically when end-to-end deep learning approaches came along in 2016. Google's WaveNet (2016) and Tacotron (2017) models set the basis

CLEARLY THE MAIN BENEFIT of the technology is audio accessibility in any language variant or custom voice at an exceptionally low cost which in turn creates the potential to apply it in much larger volumes of content that would otherwise not be localized at all in certain languages.

of modern deep learning approaches to text-to-speech, which for the first time led to the production of synthetic speech models that could rival human speech in terms of naturalness.

Our imaginations have always been sparked by robots that can speak like us and even show emotion. From the voice of Hal in 2001: Space Odyssey, to C-3PO in Star Wars and many others, we have expected machines to be able to speak like humans long before this became a reality. Today they do — and they make our imagination run wild with more applications we could use them in.

The film industry is already making use of such technologies. Young Luke Skywalker's voice in the last episode of season two of "The Mandalorian" during Mark Hamill's cameo appearance was completely synthesized due to the actor's age. So was Val Kilmer's voice in Top Gun: Maverick, which was recreated for him from earlier recordings, as his vocal cords were permanently damaged after a throat cancer operation. An entire film, Salt, has now been released, produced entirely with synthetic media, including synthetic sound and speech.

What are the properties that we look for in synthetic speech for it to be considered a viable alternative to natural speech? Naturalness refers not only to the fact that a voice needs to sound human-like, but that it sounds human-like in a specific context. Hearing a flat but otherwise perfectly human-sounding voice when we see a person yelling does not sound natural. To achieve "naturalness in context" it is not only that the technology should be able to reproduce emotions, such as anger in this example, but that aspects of it can be controlled (manually or automatically) so that the right characteristic is used at the right time in the right context.

For this to happen, the first step is a wide range of languages and dialects for which the tech is available. To create such models, tens to hundreds of hours of carefully recorded and annotated data are needed — not an easy task or immediately available in all the world's languages and dialects. The type of voice also needs to be controlled, so that it sounds like the high-pitched voice of a woman, the low-pitched voice of a man, or that of child. The same goes for speaking style, such as whispering, yelling, speaking like a cartoon and so on.

Controllability of the speech rate, i.e., speed, is an essential parameter for applications that involve time constraints, such as revoicing in video localization. On top of this, most types of revoicing require emotion, so it is also important to be able to control this too, so that a voice can sound happy or sad, or excited as needed, or display any of the range of emotions humans can express with their voice. It is the lack of emotion that makes us label a voice as robotic, or synthetic, even if it sounds perfectly natural otherwise.

Another advanced topic in text-to-speech research is adaptive speech synthesis, or in other words the ability to recreate a target voice that takes on the characteristics of a source voice one is trying to mimic, in terms of pitch, accent, pace, emotion and so on. The latest advancements in the field are closing the gap towards zero-shot speaker adaptation with great success, which makes one think of applications such as live automatic revoicing of news bulletins or any type of live event.

Clearly the main benefit of the technology is audio accessibility in any language variant or custom voice at an exceptionally low cost which in turn creates the potential to apply it in much larger volumes of content that would otherwise not be localized at all in certain languages. This is certainly the case for audio described content, a service typically addressed to a sight-impaired audience, who has benefited from the use of synthetic voices even in the pre-deep-learning era. Now it is possible not only to use AI to voice an audio description script, but to do so in an emotional voice if you so prefer, instead of a neutral voice that has been the case until now.

The quality of the technology today has reached a level that opens a new world of opportunities. Auto-narration by an artificial voice is a service that publishers already use in some cases, which recently caused a stir in the

Dr. Volker Steinbiss *is the managing director at AppTek GmbH and staff member at RWTH Aachen University. He holds a PhD in mathematics and worked on speech recognition at Philips Research in the late 1980s, before his interests broadened into more fields of human language technology, such as speech translation, synthetic speech, and natural language processing.*
vsteinbiss@apptek.com @AppTek_McLean

audiobook arena. One of the strongest arguments in favor of auto-narration is the possibility it offers to convert into audio a great list of titles that would otherwise have never been voiced. Earlier this year, a Chinese video game producer replaced the voice of a popular game character with a synthetic version of his voice when the actor was no longer available to go to the recording studio.

Any content owner could have their own bespoke voice created, for use in their content with no issues over ownership and IP rights or risk of a contractual period ending. Synthetic voices have the potential to become famous brand voices. What it takes, for now at least, to make them indistinguishable in context to human voices is a 'proof listener' who will fine-tune the output of the machine, its pace, pitch, intonation, emotion, etc. via an interface that facilitates such synthetic speech editing, until the machine performance is sculpted to perfection.

Similarly to how audio experts create sound effects, could the ability to give birth to performances out of synthetic speech "actors" result in "synthetic speech director" being one of the new jobs that will emerge in a brave new era of storytelling? ⊞

Transforming Production Work Flows through Multilingual AI Technologies

Scan the QR Code to get a glimpse into the latest advancments in AI-enabled automatic dubbing technologies.

AppTek

TRANSCRIBE. TRANSLATE. SYNTHESIZE.

WWW.APPTEK.COM