36
CLINICAL APPLICATIONS OF SPEECH SYNTHESIS

Martine Smith and John Costello

Introduction

Roughly 0.5% of the population have such difficulties producing speech that is intelligible to others that they require some way of augmenting or replacing their natural speech (Creer, Enderby, Judge, & John, 2016). This group includes children with developmental conditions such as cerebral palsy, who may never produce speech that can be understood even by their parents; it also includes young people and adults who lose the ability to speak, sometimes abruptly (e.g. post brain injury) and sometimes gradually due to neurodegenerative illnesses such as Motor Neurone Disease (MND)/Amyotrophic Lateral Sclerosis (ALS). For all these individuals, synthetic speech offers a potential opportunity to participate as a speaker in social interactions, to gain or regain a voice and to represent themselves as individuals in their speech community. Over recent years, synthetic speech has found its way into routine technologies, including navigation systems, gaming platforms, utility company service supports and a host of other industries. It is easy to assume that replacing natural speech with synthetic speech should be straightforward, but such an assumption underestimates the complexity of the contribution that natural spoken voice makes to communication, as well as the role it plays in identity formation and social participation.

The production of human speech is both complex and rapid. The speaker must decide what to say, how to say it and modulate that *how* in order to achieve a specific communicative goal. Each speaker generates a unique and personal voice, and each message uttered carries features specific to that moment of production. Although the linguistic content of the message spoken is represented by the articulated form spoken, the communicative intent of that message is carried in other acoustic features of the speech signal, as well as through aspects of nonverbal communication. The same sounds or words uttered with different intonation, pacing or stress can convey quite different intents. Given the complexity of these multiple layers of activity, replicating these processes using technology has presented many challenges.

Individuals with severe impairments often need to augment their speech using a broad range of strategies and tools, all of which come under the umbrella term of augmentative and alternative communication (AAC). Within these tools, devices that produce synthetic speech (variously termed Voice Output Communication Aids, VOCAs, or Speech Generating Devices, SGDs) can play an important role. The rest of this chapter considers current clinical applications of speech synthesis, recognizing that high-tech communication devices represent only one component of

rich multi-modal communication systems for individuals who use AAC. We start with some key concepts and a very brief review of how speech synthesis has evolved over the past decades.

Terminology and the historical context

Digitized speech is a recorded sample of natural speech that is transformed into a digitized format. Once recorded, content can be stored, replayed, segmented and reconfigured. If high quality recording equipment is used, digitized speech can be almost indistinguishable acoustically from natural speech. On the other hand, what can be replayed (i.e. "spoken") is restricted to the content that has been recorded, making unplanned, spontaneous comments impossible. Synthetic speech, by contrast, is computer-generated, using algorithms. Although some very early "speech machines" produced at the end of the eighteenth century could produce single phonemes, the VODER presented by Homer Dudley in 1939 is often regarded as the first synthesizer capable of producing continuous human speech electronically (Delić et al., 2019). Speech sounds were generated using a special keyboard, essentially the basis of what is known as a text-to-speech (TTS) procedure, now ubiquitous in speech synthesis. The advantage of a TTS system is that, unlike with digitized speech, there are no constraints on what can be generated in text and therefore what can be "spoken." However, a disadvantage is that text is composed of distinct units—letters and words with discrete boundaries, unlike the fluid acoustic wave form of continuous speech. As a consequence, early speech synthesizers sounded very robotic and unnatural. Much of the early work in speech synthesis focused on achieving a more human-sounding speech output. The emergence of Dennis Klatt's DECTalk synthesizer in 1976 heralded a new era of speech technologies. DECTalk was pioneering in that it incorporated sentence-level phonology into the algorithms used to generate speech (e.g. Klatt, 1987), thereby capturing some of the features of coarticulation so fundamental in speech production and it offered a range of different voices (male, female and child versions; see https://acousticstoday.org/klatts-speech-synthesis-d/ for some recorded examples of these early voices). From the outset, Klatt was interested in the application of this technology to support those with disabilities. The most widely-used DECTalk voice, *Perfect Paul*, based on Klatt's own voice, became best known as the voice of Stephen Hawking, the renowned theoretical physicist who died in 2018.

Since the emergence of DECTalk, both the technology underpinning speech synthesis as well as the applications in which it is now embedded have changed fundamentally. Boundaries between assistive and mainstream technology have blurred. The divisions between pre-recorded digitized speech and text-to-speech synthesis have shifted. Most modern speech synthesizers are based on extensive recorded samples of natural speech. The range of synthetic voices available has expanded greatly through advances in capturing and reconfiguring natural voice data (Pullin & Hennig, 2015). Ironically, the increasingly natural-sounding synthetic voices have led to new concerns about involuntary voice capture, ethics and cyber security (Bendel, 2019; Delić et al., 2019). Despite such concerns, these advances have brought significant benefits to people with severe speech impairments.

Clinical considerations: Speech synthesis for communication

Synthetic speech and intelligibility

Clearly for any speech synthesis to be useful for functional communication, it must be intelligible to interaction partners. Historically, synthetic speech has been found to be consistently less intelligible than natural speech (Giannouli & Banou, 2019), but over recent decades there

has been considerable progress in enhancing the quality of the acoustic signal. Manipulating speech rate, fundamental frequency variation and the overall duration of the acoustic wave (i.e. single words versus sentences) can all impact perceived intelligibility (Alamsaputra, Kohnert, Munson, & Reichle, 2006; Vojtech, Noordzij, Cler, & Stepp, 2019). However, intelligibility is not just a function of the speech signal: signal-independent variables, such as background noise, context and the "ear of the listener" (Beukelman & Mirenda, 2013) are equally important. Among native speakers of English, variables such as linguistic redundancy within the message generated, visual display of key words, contextual information and practice or familiarity with synthesized speech have all been found to enhance performance on intelligibility tasks (e.g. Alamsaputra et al., 2006; Koul & Clapsaddle, 2006). The language competence of the listener (particularly receptive vocabulary) is another important signal-independent variable. Young children, those with additional disabilities, and non-native speakers all perform less well on tasks involving synthetic speech (Alamsaputra et al., 2006; Drager & Reichle, 2001; Drager, Clark-Serpentine, Johnson, & Roeser, 2006). These groups are also more impacted by background noise, the complexity of the intelligibility task and extraneous distractions. These differences are not insignificant. Young children acquiring language using SGDs may abstract less information from the acoustic signal of synthetic speech than adults, making it more difficult for them to determine if the message generated matches the intended target and providing them with less rich phonological and phonetic information. They may also be more vulnerable to the impact of background noise in a busy classroom when they are using their SGDs. This difference may partly explain the inconsistent findings on the benefits of providing speech output in supporting literacy skills and in enhancing graphic symbol acquisition (Koul & Schlosser, 2004; Schlosser & Blischak, 2004). The low-paid status of many caring roles means that personal assistants employed to support people who rely on AAC are often migrant workers, for whom the intelligibility of synthetic speech may exacerbate other communication barriers related to linguistic competence.

Prosody and naturalness in synthetic speech

Even if synthesized speech is intelligible, inferring what others really *mean* by what they say involves more than decoding the words that have been uttered. Communicative intent is largely expressed through vocal tone, volume, intonation and stress. Natural voice can convey a speaker's personal state, regardless of the words spoken (think of the multiple meanings of "*I'm fine*"), and can even reveal information on health status (Podesva & Callier, 2015). Capturing these features using synthetic speech is particularly challenging in spontaneous utterances, constructed in-the-moment, in response to a specific communicative situation. Prosody allows speech to be intelligible, but also to sound natural. Without access to intonation or stress, individuals who rely on SGDs cannot easily express irony, humor, sarcasm, frustration or sadness. For natural speakers, calling someone's name in a particular tone of voice simultaneously attracts attention and gives an indication as to whether simply responding "*I'm here*" is sufficient or whether urgent assistance is needed (see also Pullin & Hennig, 2015). Without access to prosodic cues, these distinctions must be expressed through words, requiring far greater effort of the speaker – but also of the listener, who must interpret what has been said. Most work on incorporating prosody into synthetic speech involves generating algorithms based on linguistic rules or inferences about phrase structure from a source text, or through voice conversion, where linguistic content is mapped onto target spectral and prosodic features of a given natural voice (Vojtech, Noordzij, Cler, & Stepp, 2019). However, such systems work optimally when there is adequate grammatical information available, unlike the output of many

individuals who use aided communication (e.g. von Tetzchner, 2018). While it may be possible to develop an interface that allows a user to directly control prosody, explicitly manipulating prosody may come at a cost of user effort, in what is already an exceptionally effortful communication process.

Speech naturalness allows listeners to focus on the meaning of the message, but also impacts on the extent to which they are willing to engage with a speaker. In general, listeners have been found to prefer listening to natural speech over synthesized speech, but if natural speech is dysarthric (as is the case for many people who rely on SGDs), they may switch preference to synthesized speech (Stern, Chobany, Patel, & Tressler, 2014). Such preferences are complicated by factors such as the listener's familiarity with the speaker, the nature of the interaction and even simply knowing that the speaker has a disability. Listeners are predisposed to make judgements about a speaker based on vocal characteristics and so vocal quality is important not only for communication but also for social identity.

Clinical considerations: Voice, identity and speech synthesis

Each individual's voice represents their unique acoustic fingerprint (Costello, 2000) and a core part of their identity (Nathanson, 2017). That aspect of identity is important to self, but is also important in building and maintaining social connections (McGettigan, 2015). Familiar voices have a unique impact on listeners who have a personal relationship with that speaker (Sidtis & Kreiman, 2012), capable of instantly evoking a whole range of emotions. Hearing a familiar voice can impact stress levels (as measured by levels of hydrocortisol) in much the same way as being in the physical presence of that person (Seltzer, Prososki, Ziegler, & Pollak, 2012).

Children who have never had the ability to produce natural speech themselves risk missing out on this core aspect of identity formation – hearing their own personal voice – as well as experiencing the changes in the acoustic properties of that voice as a tangible marker of aging. Adults who lose the ability to produce intelligible speech face a potential associated loss of identity, while their loved ones face the risk of losing this critical dimension of that person, and with it the loss of personal connection over distance. These distinctions between developmental and acquired conditions create particular sets of expectations of what a synthetic voice might offer and enable, as well as impacting the willingness of individuals to accept a voice that is not their "own."

The role of personal voice in identity raises a number of questions for individuals who rely on synthetic speech. Can a synthetic voice become "personal"? If so, what are the implications if that synthetic voice is used by others, becomes obsolete or is no longer available due to technological developments? In the 80s and early 90s, the choice of synthetic voices available for communication aids was extremely limited. Many individuals with speech impairments opted to use the same voice, regardless of their age or gender. This preference often resulted in unusual interactions where several speakers (male and female) sounded exactly the same. In one day center familiar to the first author, a blind service user became adept at identifying speakers by the pacing of speech output and the sound of switch activations unique to that speaker. Despite the fact that they were not unique owners of a particular synthetic voice, some individuals (including Stephen Hawking) became attached to that voice, perceiving it as part of their identity, and rejecting offers to upgrade to more advanced, natural voices.

For those who can use natural speech, changes in voice quality and range are most rapid and pronounced in childhood, remaining relatively stable across early and middle adulthood, with very gradual, even imperceptible change in older age (Stathopoulos, Huber, & Sussman, 2011). This pattern is in marked contrast to the step-wise changes in synthetic speech that have characterized the past decades and the abrupt and often unplanned transitions in voice that are

often associated with upgrading or replacing SGDs. Individuals may start out using a child voice and in adolescence face explicit choices about transitions to a new voice that is unfamiliar to those in their environment, who may not "recognize" their voice. Furthermore, despite technical advances, available voice types continue to largely reflect gender norms with US or British standardized accents for English-language synthesizers. Alan Martin, a long-time user of an SGD commented, "Although I now have an English male voice, I sound more like a BBC news reader than the 'Scouser' that is the real ME" (Martin & Newell, 2013, p. 99) (see https://www.youtube.com/watch?v=xsqInns6LXQ where Lee Ridley comments on the familiarity of his voice because of the listeners' experience of automated train timetable announcements). The aspiration of capturing the unique nuances of individual voices so that individuals can express their identity is still a work in progress.

Voice banking, message banking, and double dipping

Two innovations in speech technology have brought the holy grail of truly personalized voices closer: voice banking and message banking. Voice banking involves using a large number of high-quality digital recordings of an individual's speech and segmenting those recordings to generate a unique phonetic database. These phonetic segments can then be combined in novel sequences to create synthetic utterances with personalized synthesized voice that can be accessed through text-to-speech (Bunnell & Pennington, 2010; Pullin, Treviranus, Patel, & Higginbotham, 2017). To complete this process, individuals register with a voice banking service and using a computer with internet connection, record in a quiet environment with a high-quality microphone. Up to 1600 phrases may be needed to establish a sufficiently large data set to create a personal voice. More recently, some voice banking services have applied Deep Neural Network (DNN) technology to this process; for example, Acapela® allows for a personalized voice to be created with only 50 recordings. VocalID™ (www.vocalid.com) was an early innovator in creating unique synthesized voices by combining the residual vocalization of an individual potential user and the recordings of a matched-speech donor from an extensive database (VocalID's Human Voicebank™; www.vocalid.com). This database was generated by volunteer donations of over 20,000 speakers representing a diversity of age, gender, cultures and geography (Patel & Threats, 2016; Pullin et al., 2017). More recent initiatives include SpeakUnique (www.speakunique.co.uk) where users can opt to bank their own voice, even if it is mildly dysarthric, or generate a bespoke voice, based on voices from an existing database or donors they nominate themselves.

Despite significant reductions in the quantity of recordings now required to create a bespoke personalized voice, individuals already experiencing speech deterioration may still find the demands of recording challenging, and there can be significant cost implications with some service providers. Voice banking is an important development in recognizing the unique and specific role of personal voice in maintaining identity; however, apart from cost and the physical demands of recording, even high-quality bespoke voices lack the facility to generate the intonation and prosody that characterizes natural speech and that reflects an individual's way of speaking.

In Message banking™ (Costello, 2000) the goal is not to generate a phonetically balanced recorded sample of speech to re-construct, but rather to capture recordings of phrases, sayings and utterances that reflect the identity and personality of the speaker. Because the emphasis is on messages rather than on words, aspects such as intonation, phrasing and emotional content are key. There are no specific constraints on the quantity of messages needed, as the focus is not on deconstructing and reconstructing the messages. Each recorded message is stored in a specific location on a speech-generating device (SGD). Message banking is a collaborative

process involving individuals at risk of losing their speech, their loved ones and their clinician. The goal is to capture that person's unique ways of speaking, their individual turns of phrase, so they can utter important messages in their own unique way (for demonstrations, see https://tinyurl.com/messagebankingprocessvideos).

Typically, individuals who use Message banking™ combine their use of these recorded messages with conventional TTS-based synthetic speech output. Double dipping™, a recent innovation pioneered by the Augmentative Communication Program at Boston Children's Hospital in collaboration with a range of partners, uses recorded messages that were completed through message banking as the script for creating a synthetic voice. The option of using one effort for both purposes (i.e. "double dipping") eliminates the need to choose which process to complete or whether to expend additional energy to complete each process separately. Successful Double dipping™ ideally requires 750 or more banked messages of sufficient quality as a database source. As one example, using their My-Own-Voice™, Acapela® uses audio files of an individual's banked messages from MyMessagebanking.com to create a personalized synthetic voice that can be fully integrated into Windows-based or iOS-based communication technology. Dozens of high-quality, synthetic voices have been created in multiple languages using 750 or more recordings from banked messages (for an Italian example, see https://youtu.be/umGQZmvRSH8).

Conclusion

There have been remarkable advances in the intelligibility and naturalness of synthetic speech since the first SGDs were introduced. Voice banking and message banking represent complementary approaches with a common purpose, that of ensuring that individuals can continue to use their own unique voice, even if vocal quality has changed due to disease progression. Both approaches recognize the importance of voice as a marker of identity and social connection, as well as a channel for communication. The next phases of innovation in speech synthesis must address the crucial role of prosody in supporting effective communication, using interfaces that are intuitive. When synthetic speech can naturally age with a user, effectively conveys emotion and communicative intent, and is uniquely configured for each individual user, all without any additional effort on the part of the user, then it will truly be approaching the power of natural speech.

References

Alamsaputra, D. M., Kohnert, K. J., Munson, B., & Reichle, J. (2006). Synthesized speech intelligibility among native speakers and non-native speakers of English. *Augmentative and Alternative Communication*, *22*(4), 258–268.

Bendel, O. (2019). The synthesization of human voices. *AI & Society*, *34*, 83–89. doi:10.1007/s00146-017-0748-x

Beukelman, D. R., & Mirenda, P. (2013). *Augmentative and Alternative Communication: Supporting children and adults with complex communication needs* (4th ed.). Baltimore MD: Brookes Publishing.

Bunnell, H., & Pennington, C. (2010). Advances in computer speech synthesis and implications for assistive technology. In J. Mullennix & S. Stern (Eds.), *Computer synthesized speech technologies: Tools for aiding impairment* (pp. 71–91). Hershey, PA: IGI Global.

Costello, J. (2000). AAC intervention in the intensive care unit: The Children's Hospital Boston model. *Augmentative and Alternative Communication*, *16*, 137–153.

Creer, S., Enderby, P., Judge, S., & John, A. (2016). Prevalence of people who could benefit from augmentative and alternative communication (AAC) in the UK: Determining the need. *International Journal of Language and Communication Disorders*, *51*, 639–653. doi:10.1111/1460-6984.12235

Delić, V., Peric, Z., Sećujski, M., Jakovljevic, N., Nikolić, N. J., Mišković, D., … Delić, T. (2019). Speech technology progress based on new machine learning paradigm. *Computational Intelligence and Neuroscience* (Article ID 4368036), 1–19. doi:10.1155/2019/4368036

Drager, K., & Reichle, J. (2001). Effects of age and divided attention on listeners' comprehension of synthesized speech. *Augmentative and Alternative Communication*, *17*(2), 109–119.

Drager, K., Clark-Serpentine, E. A., Johnson, K. E., & Roeser, J. L. (2006). Accuracy of repetition of digitized and synthesized speech for young children in background noise. *American Journal of Speech Language Pathology*, *15*(2), 155–164. doi:15/2/155 [pii]10.1044/1058-0360(2006/015)

Giannouli, V., & Banou, M. (2019). The intelligibility and comprehension of synthetic versus natural speech in dyslexic students. *Disability and Rehabilitation: Assistive Technology*. doi:10.1080/17483107.2019.1629111

Klatt, D. (1987). Review of Text-to-Speech conversion for English. *Journal of the Acoustical Society of America*, *82*(3), 737–793.

Koul, R., & Clapsaddle, K. (2006). Effects of repeated listening experiences on the perception of synthetic speech by individuals with mild-moderate intellectual disabilities. *Augmentative and Alternative Communication, 22*(2), 112–122. doi:10.1080/07434610500389116

Koul, R., & Schlosser, R. (2004.). Effects of synthetic speech output on the learning of graphic symbols of varied iconicity. *Disability and Rehabilitation*, *26*(21/22), 1278–1285.

Martin, A. J., & Newell, C. (2013). Living through a computer voice: A personal account. *Logopedics Phoniatrics Vocology*, *38*(3), 96–104. doi:10.3109/14015439.2013.809145

McGettigan, C. (2015). The social life of voices: Studying the neural bases for the expression and perception of the self and others during spoken communication. *Frontiers in Human Neuroscience*, *9*, 1–4. doi:10.3389/fnhum.2015.00129

Nathanson, E. (2017). Native voice, self-concept and the moral case for personalized voice technology. *Disability and Rehabilitation*, *39*(1), 73–81. doi:10.3109/09638288.2016.1139193

Patel, R., & Threats, T. (2016). One's voice: A central component of personal factors in augmentative and alternative communication. *Perspectives of te[illegible]HA Special Interest Groups*, *SIG 12*, 94–98. doi:10.1044/persp1.SIG12.94

Podesva, R., & Callier, P. (2015). Voice quality and identity. *Annual Review of Applied Linguistics*, *35*, 173–194.

Pullin, G., & Hennig, S. (2015). 17 ways to say 'Yes': Toward nuanced tone of voice in AAC and speech technology. *Augmentative and Alternative Communication*, *31*(2), 170–180. doi:10.3109/07434618.2015.1037930

Pullin, G., Treviranus, J., Patel, R., & Higginbotham, J. (2017). Designing interaction, voice, and inclusion in AAC research. *Augmentative and Alternative Communication*, *33*(3), 139–148. doi:10.1080/07434618.2017.1342690

Schlosser, R., & Blischak, D. (2004). Effects of speech and print feedback on spelling by children with autism. *Journal of Speech, Language and Hearing Research*, *47*(4), 848–862.

Seltzer, L., Prososki, A., Ziegler, T., & Pollak, S. (2012). Instant messages vs speech: Why we still need to hear each other. *Evolution and Human Behavior*, *33*, 42–45. doi:10.1016/j.evolhumbehav.2011.05.004

Sidtis, D., & Kreiman, J. (2012). In the beginning was the familiar voice: Personally familiar voices in evolutionary and contemporary biology of communication. *Integrative Psychological and Behavioral Science*, *46*, 146–159.

Stathopoulos, E., Huber, J. E., & Sussman, J. E. (2011). C[illegible]ges in acoustic characteristics of the voice across the lifespan: measures from individuals 4-93 years of age. *Journal of Speech Language Hearing Research*, *54*(4), 1011–1021. doi:10.1044/1092-4388

Stern, S., Chobany, C., Patel, D., & Tressler, J. (2014). Listeners' preference for computer-synthesized speech over natural speech of people with disabilities. *Rehabilitation Psychology*, *59*(3), 289–297. doi:10.1037/a0036663

Vojtech, J., Noordzij, J., Cler, G., & Stepp, C. (2019). The effects of modulating fundamental frequency and speech rate on the intelligibility, communication efficiency and perceived naturalness of synthetic speech. *American Journal of Speech-Language Pathology*, *28*, 875–886. doi:10.1044/2019_AJSLP-MSC18-18-0052

von Tetzchner, S. (2018). Introduction to the special issue on aided language processes, development, and use: An international perspective. *Augmentative and Alternative Communication*, *34*(1), 1–15. doi:10.1080/07434618.2017.1422020