A large-scale comparison of two voice synthesis techniques on intelligibility, naturalness, preferences, and attitudes toward voices banked by individuals with amyotrophic lateral...

Article *in* Augmentative and alternative communication (Baltimore, Md.: 1985) · October 2023

DOI: 10.1080/07434618.2023.2262032

CITATIONS
0

READS
58

6 authors, including:

Jolene Hyppa-Martin
University of Minnesota Duluth
14 PUBLICATIONS **129** CITATIONS
SEE PROFILE

Jason Lilley
Nemours Children's Health System
24 PUBLICATIONS **48** CITATIONS
SEE PROFILE

Mo Chen
University of Saint Joseph (Macao)
28 PUBLICATIONS **195** CITATIONS
SEE PROFILE

Abstract

Amyotrophic lateral sclerosis (ALS) commonly results in the inability to produce natural speech, making speech-generating devices (SGDs) important. Historically, synthetic voices generated by SGDs were neither unique, nor age- or dialect-appropriate, which depersonalized SGD use. Voices generated by SGDs can now be customized via voice banking and should ideally sound uniquely like the individual's natural speech, be intelligible, and elicit positive reactions from communication partners. This large-scale 2 x 2 mixed between- and within-participants design examined perceptions of 831 adult listeners regarding custom synthetic voices created for two individuals diagnosed with ALS via two synthesis systems in common clinical use (waveform concatenation and statistical parametric synthesis). The study explored relationships among synthesis system, dysarthria severity, synthetic speech intelligibility, naturalness, and preferences, and also provided a preliminary examination of attitudes regarding the custom synthetic voices. Synthetic voices generated via statistical parametric synthesis trained on deep neural networks were more intelligible, preferred, and natural than voices produced via waveform concatenation, and were associated with more positive attitudes. The custom synthetic voice created from moderately dysarthric speech was more intelligible than the voice created from mildly dysarthric speech. Clinical implications and factors that may have contributed to the relative intelligibilities are discussed.

Keywords: voice banking, synthetic speech, speech-generating device, dysarthria, amyotrophic lateral sclerosis, intelligibility, naturalness, preference, attitude, statistical parametric speech synthesis, deep neural networks, unit-selection synthesis, waveform concatenation.

Amyotrophic lateral sclerosis (ALS) is a neurodegenerative disease that progressively impairs the ability to move and breathe (ALS Association, 2022; Hanson et al., 2011). Prevalence of ALS increases around 55 years of age but can affect individuals who are decades younger (ALS Association, 2022). Many individuals survive for only three to five years after diagnosis, making post-diagnostic communication with loved ones and health care providers of great importance (ALS Association, 2022; Zhang et al., 2021). Unfortunately, a majority of individuals diagnosed with ALS experience dysarthria that is characterized by slow, slurred, hypernasal speech, short phrase length, and vocal harshness (Hanson et al., 2011; Kühnlein et al., 2008). As dysarthria progresses, most individuals with ALS completely lose the ability to produce natural speech and benefit from augmentative and alternative communication (AAC) interventions, including speech-generating devices (SGDs; Doyle & Phillips, 2001).

AAC acceptance rates tend to be high among individuals diagnosed with ALS. For example, Ball and colleagues (2004) reported on 50 individuals diagnosed with ALS and found that 96 percent used AAC to communicate until the end of their lives. Historically, the synthetic voice generated by an SGD has been one of a family of commercial voices licensed by manufacturers, which caused most individuals who communicated via SGDs to produce synthetic speech that was neither unique, nor age- or dialect-appropriate (Creer et al., 2013). This restricted variety of voices tended to depersonalize SGD use, which can be strongly felt by adults experiencing ALS who face a loss of their vocal identity and a terminal illness.

In recent years, voice banking programs such as CereProc (CereProc Limited, 2022), Acapela (Acapela Group, 2022), and ModelTalker (Nemours Children's Health, 2022) have enabled the creation of custom synthetic voices that are recognizably similar to the individual's natural speech. Voice banking is a process that involves systematically recording the natural

speech of an individual before they lose the ability to speak due to a condition like ALS (ALS Association, 2020; Veaux et al., 2011). A custom synthetic voice is generated from the recordings and loaded onto an SGD, thus preserving the individual's vocal identity and ability to communicate, even after natural speech is lost. Custom synthetic voices that sound like the individual's familiar natural speech have been viewed positively (Costello, 2018; Kraker, 2018).

There are two common processes used to create synthetic speech output: waveform concatenation synthesis (WCS) and statistical parametric speech synthesis (SPSS; Black et al., 2007; Bunnell & Pennington, 2010; Zen et al., 2009), the latter of which is subsequently referred to as statistical parametric synthesis (SPS) for brevity. Different synthesis processes tend to affect listeners' perceptions of the synthetic speech output (e.g., Ling et al., 2006). WCS is a mature technology that dominated text-to-speech systems in the late 1990's and early 2000's and has been developed and refined since then (Bunnell & Pennington, 2010; Hunt & Black, 1996; Zen et al., 2009). When used with voice banking, WCS produces novel synthetic speech output from an SGD by concatenating small units of the person's actual recorded natural speech. WCS systems can produce synthetic speech output that sounds very similar to the individual's own natural speech and voice because the output is comprised of recordings of their actual natural speech (Black et al., 2007; Bunnell & Pennington, 2010; Hunt & Black, 1996). However, WCS requires a very large number of recorded utterances to avoid unnatural prosodic variations and discontinuities (i.e., breaks or gaps) in the synthetic speech and tends to sound best when created from at least one hour of recorded speech, though substantially more speech is needed for truly natural sounding speech prosody (Bunnell & Pennington, 2010; Yamagishi et al., 2012). Capturing one hour of recorded speech requires multiple hours of recording, which can be challenging to individuals with ALS due fatigue and the time commitment involved.

By contrast, SPS systems produce synthetic speech output based on a statistical model of the individual's recorded natural speech (Zen et al., 2009). SPS has evolved from the Hidden Markov Models that were common in the early 2000's and is presently dominated by systems that use deep neural networks (DNNs; Bunnell & Pennington, 2010; Zen et al., 2009). In either case, during synthetic voice construction, the model learns a millisecond-by-millisecond mapping from linguistic features to vocoder input parameters, and in synthesis the model generates vocoder parameters from which the vocoder produces synthetic speech output.

Further advances have led to SPS voices in which the vocoder itself has also been replaced by a DNN and can produce synthetic speech that is very difficult to distinguish from natural speech (Shen et al., 2018; van den Oord et al., 2016). This is exciting for clinicians who seek to provide synthetic voices that are highly customized, natural sounding, and easy to create for their clients' SGDs. However, the computational demands of the software for these voices tends to preclude their application to SGDs at this time. Consequently, present day SGDs using an SPS system tend to generate synthetic speech that sounds more artificial and less similar to the speech of the individual who made the recordings when compared to a voice produced via WCS (Bunnell & Pennington, 2010; Capes et al., 2017; Creer et al., 2013). However, present day SGDs that use SPS produce voice output that is relatively free of discontinuities and has more natural prosody than WCS output (Bunnell, 2010; Kuligowska et al., 2018). Moreover, voices produced via SPS require fewer recordings than WCS systems (Bunnell & Pennington, 2010; Yamagishi et al., 2012) which is appealing to individuals experiencing ALS and their clinicians.

Many factors may influence a listener's preferences regarding synthetic voices (e.g, Hinterleitner, 2017), including how natural the voice sounds (Baird et al., 2018; Mayo et al., 2011); how much the synthetic voice sounds like the original speaker (Mills et al., 2014); the

personality or attitude conveyed (Aylett et al., 2017;); and voice intelligibility (Bunnell, 2010; Mills et al., 2014). The synthesis system may also influence preference. For example, among 254 individuals who recently completed the voice banking recording process and then had voices created using both WCS and SPS, 164 (65%) preferred the voice produced by an SPS system, while 90 (35%) preferred WCS (Lilley et al., 2020). Further exploration to identify factors that influence preferences about synthetic voices could contribute to making voice banking more efficient. Additionally, communication requires a successful exchange of information between a sender and a receiver. Hence, potential communication partners' perceptions about voices also warrant examination. Understanding preferences of peers and other communication partners could help identify characteristics of synthetic speech that support the acceptance of AAC and promote positive views of its use (McCarthy & Light, 2005; Veaux et al., 2011).

Obviously, synthetic speech must also be highly intelligible. Creating intelligible, custom synthetic speech is an especially important issue for individuals who already have dysarthria at the time they record their natural speech for voice banking. The intelligibility of the final custom synthetic voice can depend on the individual's speech characteristics at the time of recording (Mills et al., 2014). Individuals with ALS are encouraged to record their speech before it is affected by disease progression (Lilley et al., 2020). However, dysarthria is often an early or initial symptom of ALS (Hanson et al., 2011; Veaux et al., 2011). A sample of about 350 individuals diagnosed with ALS who recently completed voice banking showed that 54% already had dysarthria at the time of recording (Lilley et al., 2020), which may affect both intelligibility and naturalness of the resulting custom synthetic voice.

Naturalness refers to whether synthetic speech is perceived as uniquely human, despite being computer-generated (Nusbaum et al., 1995). Not surprisingly, more natural synthetic

voices have been associated with more comfortable communication exchanges (Nusbaum et al., 1995). Identifying natural-sounding synthetic voices for persons who communicate using SGDs is widely regarded as important (Mayo et al., 2011; Nusbaum et al., 1995), but measuring naturalness tends to be less straightforward than measuring intelligibility (Baird et al., 2018; Nusbaum et al., 1995). During an evaluation of synthetic speech naturalness, some listeners placed more importance on the presence of discontinuities in synthetic speech and less importance on aspects like stress and intonation (Mayo et al., 2011). Even state-of-the-art synthetic voices (e.g., Siri) are not able to fully convey all unique elements of the natural human voice, such as flexible intonation patterns or adaptation to the environment (Capes et al., 2017).

Another area of importance pertaining to SGD use is attitudes. Attitudes consist of the interaction of thoughts, feelings, and behaviors and refer to a person's psychological tendency to evaluate an entity with favor or disfavor (Eagly & Chaiken, 2007). Attitudes often predict a person's social behavior and interactions. Historically, negative attitudes have been reported toward individuals with disabilities, and numerous investigations seeking to describe attitudes toward people who communicate using AAC have been conducted (Hyppa-Martin et al., 2021; Kraus, 1995; McCarthy & Light, 2005).

Ideally, a person who communicates using an SGD would have voice output that sounds like unique dialect- and age-appropriate natural speech, is intelligible, and elicits positive reactions and attitudes from potential communication partners. Individuals diagnosed with terminal conditions like ALS need high quality AAC interventions to support effective communication with loved ones and to navigate important medical decisions. Identifying intelligible, appealing, customized AAC solutions for individuals experiencing ALS can enhance

quality of life and is an important consideration for speech scientists, speech-language pathologists (SLPs), and other interventionists.

Current options for creating custom synthetic voices include numerous WCS and SPS systems, each with unique strengths and weaknesses and the individuals banking their voices for use with these systems includes persons diagnosed with ALS who experience mild and moderate dysarthria. This study sought to extend current knowledge regarding the perceptions of potential communication partners about custom synthetic voices created for individuals with ALS and to offer insight into the effect of synthetic speech characteristics on naïve listeners' and potential communication partners' perceptions about synthetic voices. Specifically, this large-scale study compared the synthetic voice output created by two readily available real-world synthesis systems (one WCS system and one SPS system) that were used to produce synthetic voices for two individuals diagnosed with ALS, one who had mildly dysarthric speech and the other with moderately dysarthric speech. Naïve listeners completed intelligibility measures, and provided ratings of naturalness and preference. A preliminary examination of listener attitudes was also conducted. Institutional review board approval was obtained before conducting the study.

Method

Participants

Participants (Table 1) included 831 adults (324 males, 507 females) with a mean age of 44 years (*SD* = 24 years, range = 18 to 81 years). Each participant reported being fluent in English and having vision and hearing that functioned typically. Most participants reported ethnic background as White, but ethnic backgrounds also included Asian, Black/African American, Hispanic/Latino/Spanish, American Indian or Alaska Native, and Native Hawaiian or Pacific Islander. Education levels of participants ranged from no high school diploma to doctoral

degree. Participants tended to be inexperienced with AAC. More than 80% of participants (n = 670) had no training regarding communication aids, and 56% (n = 466) had never interacted with a person who used a synthetic voice to communicate.

Research Design

A 2 (dysarthria severity) × 2 (synthesis system) mixed between- and within-participants design was used. The between-participants factor was two levels of dysarthria severity observed at the time of voice banking (mild or moderate). Participants assigned to the mild condition only listened to experimental stimuli (i.e., the custom synthetic speech) created from the recordings made by the individual with mild dysarthria. Participants assigned to the moderate condition only listened to stimuli made from the recordings provided by the individual with moderate dysarthria. Assignment to condition was counterbalanced. The within-subjects factor was the synthesis system, with two levels (WCS and SPS). Thus, each participant heard stimuli from only one speaker, but both synthesis systems. The dependent variables in this study were: (a) synthetic speech intelligibility, (b) synthetic speech naturalness, (c) participants' preferences toward synthetic speech, and (d) participants' reported attitudes regarding the synthetic speech.

Materials

Survey, Hardware, and Data Collection Software. The survey had five sections. Each section was presented in the same sequence to all participants as follows: (a) 20 items that probed the naturalness of synthetic speech, (b) 10 items to measure intelligibility, (c) five items that probed preferences regarding synthetic speech, (d) eight items that probed attitudes about persons who use synthetic speech to communicate, and (e) demographic questions (Table 1). Each section is subsequently described. Each participant used a supra-aural headset, a tablet computer, and a keyboard to complete the survey. Responses were saved to a secure database.

Synthesis Systems. Two synthesis systems were selected to create the experimental voices. These systems were selected because they were readily available online via ModelTalker (Nemours Children's Health, 2022), represented systems that were actively used in real-world clinical applications to create custom synthetic voices for individuals with ALS, and were among the options offered by the first author's university voice banking clinic. The first system, MT-WCS, was a WCS system that generated synthetic speech via waveform concatenation. The second system, MT-SPS, was a DNN-based SPS system modified from Merlin (Wu et al., 2009) that used a standard parametric vocoder (Morise et al., 2016).

Synthetic Voice Stimuli. The synthetic stimuli were generated from the recorded banked voices of two adult females who consented to the use of their voices for research. Each had been diagnosed with ALS by a neurologist. One individual was diagnosed with mild dysarthria, and the other with moderate dysarthria, by a licensed SLP with extensive expertise in motor speech disorders. Both speakers' consonants were sometimes imprecise. Unlike the mildly dysarthric speaker, the moderately dysarthric speaker also had a notably slower speech rate, lower pitch, reduced loudness, shorter phrase length per inhalation, and vocal quality that was breathy, hoarse, and sometimes strained. Both were middle-aged, native speakers of American English, and recorded the same 1600 sentences in a controlled recording environment while following established recording protocols (e.g., Hyppa-Martin et al., 2017; Westley et al., 2019). Each corpus of recordings (i.e., one corpus of mildly dysarthric speech, the other of moderately dysarthric speech) was used to create two synthetic voices, one voice that was produced via MT-WCS, and the other produced via MT-SPS. This process resulted in four experimental synthetic voices: (a) a WCS voice created from mildly dysarthric speech, (b) an SPS voice created from mildly dysarthric speech, (c) a WCS voice created from moderately dysarthric speech, and (d) an

SPS voice created from moderately dysarthric speech. The synthetic voice stimuli were presented at the same intensity during the experiment with root mean square amplitude of each stimulus phrase at 70 dB.

Setting

Participants were recruited and data were collected for three days at a university research facility located on state fair grounds. The fair was attended by thousands of people who were notified of the opportunity to participate via advertisements posted on fairgrounds and social media. The research facility had several separate areas (bays) where data collection occurred.

Procedures

Recruitment and Incentives. Potential participants approached the bay, confirmed eligibility, sat at one of nine stations, placed sanitized headsets over their ears, and self-administered the electronic survey, which began with a consent form. Upon completion, participants spun a prize wheel and received a prize such as a book bag or pencil.

Pre-test. Prior to beginning the survey, each participant listened to an audio file of synthetic speech unrelated to the experimental stimuli. The pre-test instructed participants to adjust headphone loudness to the desired level and ensured that participants had adequate vision and fine motor skills to navigate the tablet's touchscreen.

Intelligibility. Each participant listened to 10 syntactically correct, but semantically unpredictable sentences (SUSs; Benoit et al., 1996). An example of an SUS is, "Take the smart pool behind a potato." SUSs use familiar vocabulary, but place it in a sequence that cannot be easily predicted, preventing participants from using real-world linguistic knowledge to decipher the sentence. The SUSs were created for this study using a custom software algorithm consisting of a syntactically correct sentence frame template into which randomly selected, common words

of the proper part of speech, inflection, and verb transitivity were inserted. Ten different sentence frames and a list of 1236 common English words of 1 to 3 syllables were used to generate 10 sentences from each frame, such that no nouns, verbs, prepositions, or adjectives were used more than once. In each set of 10 SUSs, five were produced by the WCS system, and five were produced by the SPS system, presented in random order. Participants were prompted to listen carefully to each spoken SUS, which played only once. After each SUS was spoken by the appropriate synthetic voice, the participant was prompted to type what they thought they had heard being spoken by the synthetic voice.

Naturalness Ratings. Each participant listened to 20 Harvard sentences (IEEE Subcommittee, 1969) produced by the experimental voice for their assigned condition (mild or moderate). For each condition, ten sentences were created using the WCS system, and the other 10 using the SPS system, presented in random order. Harvard sentences contain vocabulary familiar to most English speakers, and each set of 10 sentences presents phonemes at a similar frequency to spoken English (IEEE Subcommittee, 1969). An example is, "Ten pins were set in order." Participants listened to each sentence once, spoken by the appropriate synthetic voice, and then responded to the prompt "The speech in this sentence sounded____" by selecting an answer from a 1 to 9 scale in which 1 was "Very Unnatural," and 9 was "Very Natural."

Preference Ratings. Participants first answered four questions regarding preferences about synthetic voices, and then listened to one voice (e.g., the voice generated by the WCS system) speaking a 51-word portion of the *Rainbow Passage* (Fairbanks, 1960), followed by the other voice (e.g., generated by the SPS system) speaking the same passage. Presentation sequence was counterbalanced. Participants listened to both passages, then indicated preference.

Reported Attitudes. To preliminarily examine attitudes, participants listened to one voice (e.g., the voice generated by the SPS system for their assigned experimental condition) speaking a 132-word excerpt from the passage *Comma Gets a Cure* (Honorof et al., 2000). Next, participants answered four survey items extracted from the *Attitudes Toward Nonspeaking Persons Scale* (Gorenflo & Gorenflo, 1991). The four items were selected because they probed participants' thoughts, feelings, and emotions about a person who might use the synthetic voice to communicate, including judgments of the person's confidence, feelings of pity toward the person, and comfort in talking to the person. For example, one item was "I do not feel sorry for this person." Participants answered on a five-point scale (strongly agree, agree, undecided, disagree, strongly disagree). Next, participants heard the same passage spoken by the other experimental voice (e.g., the voice produced via the WCS system), and answered the same four survey items. Presentation sequence of SPS- and WCS-produced speech was counterbalanced, and items were scored using established procedures (Gorenflo & Gorenflo, 1991).

Demographic Items. Finally, participants responded to demographic questions (Table 1).

Data Analysis

All responses were downloaded from the secure database and analyzed using RStudio (Version 1.2.1335). The survey was piloted prior to use for experimental data collection; however, a coding error that did not become evident until hundreds of participants had completed the study affected the presentation of one group of SUSs (used to measure intelligibility) and one group of Harvard sentences (used to obtain naturalness ratings). Given the large sample size of this study, if a participant had incomplete data for one dependent variable (e.g., partial data for intelligibility), all data for that dependent variable from that participant were omitted from the analysis. However, if that same participant had complete data for another

dependent variable (e.g., complete data for naturalness), all data for that dependent variable from that participant were included in the analysis. As a result, while all 831 participants answered the demographic questions and completed preference (n = 831) and attitude ratings (n = 831), a smaller sample was obtained for intelligibility (n = 773) and naturalness analyses (n = 758).

Intelligibility. A phonetic edit distance (PED; Bunnell & Lilley, 2007) was first computed between the spoken SUS stimulus and the typed participant response provided by each of the 773 participants. PED is the minimum number of edits (e.g., phoneme insertions, deletions, or substitutions) needed to transform the phonetic transcription of the participant response into the stimulus transcription. A more intelligible stimulus will have a lower PED. Next, PED was divided by the number of phonemes in the sentence to normalize for sentence length, resulting in a mean normalized edit distance (NED). A mixed ANOVA was used to examine the relationship among synthesis system, dysarthria severity, and intelligibility.

To compute the PED, all responses were manually inspected. All null responses and responses judged to consist solely of comments (e.g., *No idea what was said*) were scored as unintelligible. Individual response tokens that were judged to indicate unintelligible words were scored as unintelligible, leaving the rest of the response intact. Examples included English words (e.g., *unsure*), strings of random characters (e.g., *xxx*), and multiple punctuations (e.g., *??*). All other punctuation was deleted. In some cases, listeners mixed responses with commentary (e.g., *I heard shop and people but not sure what else was said*). In such cases, the comments were deleted, leaving response words (e.g., *shop people*). In a few cases, listeners indicated multiple transcriptions for a stimulus. Only the first transcription was analyzed.

Naturalness Ratings. Data from 758 participants who completed the naturalness ratings for all 20 sentences were included in the analysis. A mixed ANOVA was conducted to examine the relationships among synthesis type, dysarthria severity, and perceived naturalness.

Preference Ratings. Preferences were provided by 831 participants. Mulltiple choice items were analyzed via a chi-square test, scaled items were analyzed via *t*-test.

Reported Attitudes. Data from 831 participants who completed all four attitude items were coded according to established procedures (Gorenflo & Gorenflo, 1991). A mixed ANOVA was conducted to examine the influence of the synthesis type on reported attitudes, item by item.

Demographic Items. Demographic items were analyzed using descriptive statistics.

Results

Intelligibility

The intelligibility portion of the study was completed by 773 participants (388 participants in the mild dysarthria condition, 385 in the moderate dysarthria condition). Synthetic voices produced via the SPS system (*mean NED* = .29, *SD* = .27) were more intelligible than those produced via the WCS system (*mean NED* = .32, *SD* = .27, $t(7728) = 4.95$, $p < .001$). Surprisingly, the synthetic voices created from the recorded natural speech produced by the individual with moderate dysarthria (*mean NED* = .27, *SD* = .27) were more intelligible than the synthetic voices created from the recorded natural speech of the individual with mild dysarthria (*mean NED* = .33, *SD* = .27) $t(7728) = 9.19$, $p < .001$. The mixed ANOVA analysis (Type 3) showed no interaction effect between dysarthria severity and synthesis system ($F(1, 771) = .18$, $p = .07$, ges = .00); a significant effect of dysarthria severity ($F(1, 771) = 19.89$, $p = .00$, ges = .02); and a significant effect of synthesis type ($F(1, 771) = 34.91$, $p = .00$, ges = .006).

Naturalness

Naturalness ratings were completed by 758 participants (390 participants in the mild dysarthria condition, 368 participants in the moderate dysarthria condition). Regardless of dysarthria severity, speech produced via the SPS system was perceived as more natural than speech produced by the WCS system. With prior assumptions for homogeneity of variance being met by Levene's Test ($F(3, 15156) = .74$, $p = .53$), the mixed ANOVA (Type 3) showed no interaction effect between dysarthria severity and synthesis system ($F(1, 756) = 3.37$, $p = .07$, ges = .00); no effect of dysarthria severity ($F(1, 756) = 2.07$, $p = .15$, ges = .00); and a significant effect of synthesis system ($F(1, 756) = 537.50$, $p = .00$, ges = .08). Participants in the mild dysarthria condition rated speech produced by MT-SPS an average of 5.21 ($SD = 2.09$, range = 1-9), and speech produced by the MT-WCS an average of 4.49 ($SD = 2.08$, range = 1-9). Stimuli in the moderate dysarthria condition produced by MT-SPS averaged 5.40 ($SD = 2.10$, range = 1-9) and stimuli produced via MT-WCS averaged 4.56 ($SD = 2.07$, range = 1-9).

Preferences Regarding Maintaining Vocal Identity, Naturalness, and Intelligibility

To examine whether participants had preferences about maintaining vocal identity, they used a 5-point scale (strongly agree, agree, undecided, disagree, strongly disagree) to indicate whether they preferred to use a synthetic voice that sounds "as much like their own voice as possible" should they need to use one for the rest of their life. Regardless of being in the mild or moderate dysarthria condition, participants tended to agree or strongly agree with using a synthetic voice that sounds as much like their "own natural voice as possible" if they had to communicate using synthetic speech. Ratings in the mild dysarthria condition ($n = 426$) averaged 1.66 ($SD = .79$, range = 1-5); ratings in the moderate dysarthria condition ($n = 405$) averaged 1.64 ($SD = .78$, range = 1-5); and a Welch two-sample t-test revealed no difference between conditions ($t(828.11) = .41$, $p = .68$). Similarly, most participants agreed or strongly agreed that

a person they spend a lot of time with use a synthetic voice that sounded like that person's own natural voice; condition had no effect ($t(824.69) = -.20$, $p = .85$). Ratings from participants in the mild dysarthria condition ($n = 426$) averaged 1.8 ($SD = .87$, range = 1-5) and in the moderate condition ($n = 405$) averaged 1.81 ($SD = .88$, range = 1-5).

To examine relative preferences for intelligibility and naturalness, participants were asked whether they would prefer to use a synthetic voice that sounds (a) "*different* from [their] natural voice and is almost *always* understood," (b) "*similar* to [their] natural voice and is *frequently* understood," or (c) "*exactly like* [their] natural voice and is *sometimes* understood." Most participants preferred (b), a voice that as *similar* to their natural voice and *frequently* understood by others, regardless of whether they were in the mild or moderate dysarthria condition. In the mild dysarthria condition ($n = 426$), 151 participants preferred option (a); 249 preferred option (b); 26 participants preferred option (c). Participants in the moderate dysarthria condition ($n = 405$), responded as follows: 144 participants preferred option (a), 233 participants preferred option (b), and 28 participants preferred option (c), respectively. There was no difference between conditions ($\chi(2) = .24$, $p = .89$).

To examine relative preferences for intelligibility and naturalness when listening to a synthetic voice used by somebody else, participants were asked whether they would prefer somebody they spend a lot of time with to use a synthetic voice that sounds (a) "*different* from [that person's] natural voice and is almost *always* understood," (b) "*similar* to [that person's] natural voice and is *frequently* understood," or (c) "*exactly like* [that person's] natural voice and is *sometimes* understood." Most participants preferred (b), a voice that was *similar* to that person's natural voice and *frequently* understood. In the mild dysarthria condition ($n = 426$), 131 participants preferred (a); 263 participants preferred (b); and 32 participants preferred (c). In the

moderate dysarthria condition these responses were given by 138, 237, and 30 participants, respectively. A chi-square test indicated no difference between conditions ($\chi(2) = 1.07$, $p = .59$).

Preferences for Synthesis Systems

For the 426 participants who were in the mild dysarthria condition, a chi-square test showed that preference toward a synthesis system was associated with presentation order ($\chi^2(1) = 58.70$, $p < .001$). Participants who listened to the passage produced by the WCS system first were more likely to prefer the voice produced via the SPS system. Among the 213 participants who listened to the MT-WCS voice first, 181 (85%) preferred the MT-SPS voice while 32 preferred the MT-WCS voice; among the 213 participants who listened to the MT-SPS voice first, 107 (50%) preferred it while 106 preferred the MT-WCS voice. Among the 405 participants in the moderate dysarthria condition, a chi-square test again showed that participants' preference for the voice produced by the SPS or WCS systems was associated with presentation order ($\chi^2(1) = 51.35$, $p < .001$). Of the 217 participants who listened to the MT-WCS output first, 186 (86%) preferred the MT-SPS voice while 31 preferred the MT-WCS voice. Among 188 participants who listened to the MT-SPS voice first, 100 (53%) preferred it; 88 preferred the MT-WCS voice.

Effects of Dysarthria Severity and Synthesis Type on Attitudes

The four attitude items were completed by 831 participants (426 participants in the mild dysarthria condition, 405 in the moderate dysarthria condition). For each item, prior assumptions for a mixed ANOVA analysis (Type 3) were met per the results of Levene's Test for homogeneity of variance. The items probed feelings of pity toward, willingness and comfort to talk with, and judgments of self-confidence about the person who used the synthetic voice to communicate, via a scale from 1 (most negative) to 5 (most positive). The SPS system was associated with more positive attitudes than the WCS system, regardless of dysarthria severity.

For the item: *I do not feel sorry for this person*, following Levene's Test ($F(3, 1658) = 1.96$, $p = .12$), the ANOVA showed no interaction effects between dysarthria severity and synthesis system ($F(1, 829) = .72$, $p = .40$, ges < .001); no effect of dysarthria severity ($F(1, 829) = .37$, $p = .54$, ges < .001); and a significant effect of synthesis system on attitudes ($F(1, 829) = 5.55$, $p = .02$, ges = .001). Attitudes regarding the voice created from the mildly dysarthric recordings produced via MT-SPS system averaged 3.08 ($SD = 1.15$, range = 1-5), which were slightly more positive than the MT-WCS voice from the same recordings ($M = 2.96$, $SD = 1.15$, range = 1-5). The same relationship existed between the synthetic voices created from the moderately dysarthric speech produced via MT-SPS ($M = 3.09$, $SD = 1.24$, range = 1-5) and MT-WCS ($M = 3.04$, $SD = 1.21$, range = 1-5).

For item: *This person is as self-confident as other people*, following Levene's Test ($F(3, 1658) = .80$, $p = .49$), the ANOVA revealed no interaction effect between dysarthria severity and synthesis system ($F(1, 829) = .14$, $p = .71$, ges < .001); no effect of dysarthria severity ($F(1, 829) = 2.66$, $p = .10$, ges = .002); and a significant effect of synthesis system ($F(1, 829) = 6.10$, $p = .01$, ges = .002). Regardless of the dysarthria severity, participants reported more positive attitudes toward the voice produced by the SPS system than the WCS system. The voice that used mildly dysarthric recordings with MT-SPS averaged 3.76 ($SD = .87$, range = 1-5), while its MT-WCS version averaged 3.67 ($SD = .90$, range = 1-5). The voice produced from the moderately dysarthric speech recordings via MT-SPS averaged 3.83 ($SD = .82$, range = 2-5), and the MT-WCS version averaged 3.77 ($SD = .89$, range=1-5).

For item: *I would prefer not to talk with this person*, following Levene's Test ($F(3, 1658) = .22$, $p = .88$), the ANOVA showed no interaction effect between dysarthria severity and synthesis system ($F(1, 829) = .01$, $p = .91$, ges < .001); no effect of dysarthria severity ($F(1, 829)$

= .38, p = .54, ges < .001); and a significant effect of synthesis system (F(1, 829) = 16.82, p < .001, ges = .004). The voice produced via the SPS system was associated with more positive attitudes than the WCS system. The voice created from mildly dysarthric speech via MT-SPS averaged 4.21 (SD = .89, range = 1-5) and its MT-WCS version averaged 4.10 (SD = .91, range = 1-5). The voice created from moderately dysarthric speech via MT-SPS averaged 4.24 (SD = .83, range = 1-5), which was higher than its MT-WCS version (M = 4.13, SD = .91, range = 1-5).

The item: *I would feel uncomfortable answering questions asked by this person* also indicated more positive attitudes toward the voices created with the SPS system than the WCS system, which was more pronounced among participants in the mild dysarthria condition. Following Levene's Test (F(3, 1658) = .58, p = .63), the ANOVA results revealed an interaction effect between dysarthria severity and synthesis system (F(1, 829) = 8.00, p = .005, ges = .003). No effect of dysarthria severity was detected (F(1, 829) = .05, p = .82, ges < .001). The effect of synthesis system was significant (F(1, 829) = 9.08, p = .003, ges = .003). The rating for the voice produced using mildly dysarthric speech recordings via MT-SPS averaged 3.87 (SD = 1.10, range = 1-5), which was higher than its MT-WCS version (M = 3.63, SD = 1,15, range = 1-5). The voice produced using moderately dysarthric speech recordings and MT-SPS averaged 3.77 (SD = 1.14, range = 1-5), slightly higher than MT-WCS (M = 3.76, SD = 1.08, range = 1-5).

Discussion

Many individuals experience dysarthria prior to ALS diagnosis and prior to voice banking (Hanson et al., 2011; Veaux et al., 2011) making it important to consider the interplay between dysarthria severity, synthesis systems commonly available for use with SGDs, and perceptions of potential communication partners. This study examined whether intelligibility, naturalness, preferences, and attitudes varied as a function of synthesis system when creating

custom synthetic voices from mildly and moderately dysarthric speech. Compared to prior investigations (e.g., Mayo et al., 2011; Nusbaum et al., 1995), this study recruited a substantially larger and more diverse sample of naïve listener participants. The speech recordings used in the study were representative of speech that is routinely banked by individuals with ALS, and the synthesis systems were representative of technology in real-world clinical use.

Intelligibility and Naturalness

Prior examinations have associated SPS systems with intelligible synthetic speech (Bunnell et al., 2017; Capes et al., 2017; Mills et al., 2014), and the SPS output was based on (rather than concatenated from) the dysarthric speech produced by the two women with ALS who made the recordings for this study. In addition, SPS systems are associated with synthetic voice output that has more natural prosody and fewer discontinuities than WCS systems (Bunnell, 2010; Kuligowska et al., 2018). Hence, it is not surprising that the SPS system resulted in custom synthetic voice output that was more intelligible and perceived as more natural than the WCS system, regardless of dysarthria severity.

However, it was surprising that the synthetic voices created from the moderately dysarthric speech were more intelligible than those created from the mildly dysarthric speech. Given this unexpected result, a post-hoc comparison of both women's original speech recordings was conducted to confirm the SLP's diagnosis of dysarthria severity and investigate acoustic factors of each woman's speech that may affect intelligibility (Mulligan et al., 1994; Tomik et al., 1999). First, the SLP, who had extensive experience in motor speech disorders, reassessed each woman's natural speech recordings and confirmed the initial assessment of dysarthria severity, which a second SLP also confirmed. Both SLPs noted that the moderately dysarthric speaker had a slower speech rate, shorter phrase length, and generally poorer vocal quality than

the individual with mild dysarthria. However, while both individuals with dysarthria were sometimes imprecise in producing consonants, the individual with moderate dysarthria often made effortful attempts to over-articulate consonants. By contrast, both SLPs noted that the individual judged as having mild dysarthria used a faster speech rate and had better vocal quality, but was less effortful with articulation, particularly in consonant clusters.

To investigate the SLPs' observations, additional analyses were conducted comparing the two individuals' recordings of the same 1600 sentences. To assess vowel space, measurements of F1 and F2 of the cardinal vowels /i æ ɑ u/ in non-nasal environments were made from the same utterances from each individual. Measurements from spectrograms from the temporal center of the steady states of each vowel were taken, and the resulting vowel space quadrilaterals had areas of 0.49 KHz^2 for the individual with mild dysarthria and 0.36 KHz^2 for the individual with moderate dysarthria, again supporting the notion that the mildly dysarthric speaker was less dysarthric. The individual with moderate dysarthria also had a significantly greater mean phrase duration (5883 vs. 4509 msec; $p < .001$), and produced phonemes of longer duration overall (224 msec vs. 165 msec, $p < .001$), confirming that the individual judged as having moderate dysarthria had a slower speech rate. However, there was also a clear pattern in the duration variances: using an *F* test, the individual with moderate dysarthria had significantly greater variance ($p < .01$) in mostly vocalic segments (/ɪ ɛ æ ɨ ʌ ɑ ɔ ʊ u eᴵ oᵚ aᴵ aᵚ m/), but the individual with mild dysarthria had a significantly greater variance in many obstruents (/p t d k f θ ð h/), supporting the SLPs' assessments of less precise, less consistent consonant production by the individual with mild dysarthria. Dysarthria severity is beyond the control of the individual, but these findings suggest that factors that may be within the control of individual speakers could override some aspects of dysarthria severity and affect the resulting custom synthetic voice.

Although the individual with moderate dysarthria began recording at a more dysarthric stage, her more consistent over-articulation and speaking rate, albeit slower, likely contributed to the increased intelligibility in her custom synthetic voices generated by both the WCS and SPS systems when compared to those voices created for the individual with mild dysarthria. It is possible that the individual with moderate dysarthria may have had more real-world experiences during which her communication partners did not understand her speech, and she may have begun to naturally compensate by over-articulating consonants. The individual with mild dysarthria may not have had these experiences.

These behavioral factors are important for voice bankers (and for those assisting them with voice banking) to monitor. Most healthcare facilities that provide SLP services do not provide voice banking services to individuals diagnosed with ALS, but some universities offer voice banking support (e.g., Kraker, 2018). Universities have a mission to serve their community, provide educational opportunities for students, and often have the infrastructure (e.g., sound dampening recording booths, high speed internet, technical support services) that make them ideal resources for individuals who benefit from voice banking with support. Hence, at least for some individuals, making speech recordings in a controlled setting with a support person may be a good option. In addition, such support may concurrently provide valuable learning opportunities for SLP graduate students to serve individuals who experience dysarthria by providing instruction to maximize and monitor the consistency of breath support, speech rate, and articulation. Notably, the onset of the COVID-19 pandemic prompted some universities to begin offering this type of voice banking support via telepractice which may make accessing it more convenient to individuals diagnosed with ALS. Additionally, online voice banking applications have become more user-friendly, and many individuals bank their voices from the

comfort of their own homes with no support at all. In these circumstances, it may be beneficial for voice banking software to provide automated feedback regarding rate and articulation, and to do so in a manner that is not discouraging to the person making the recordings.

Listener Preferences Regarding Naturalness, Intelligibility, and Vocal Identity

Similar to previous studies (e.g., Creer et al., 2013), participants in this study resoundingly indicated a preference to maintain one's vocal identity and use a custom synthetic voice that sounded as much like one's own natural voice as possible, regardless of whether the person who needed to use the synthetic voice was the participant or a person with whom they spend a lot of time. Interestingly, participants in the present study preferred to maintain vocal identity even if it meant sacrificing some intelligibility, and this preference was more pronounced among participants who listened to the voice produced via the MT-WCS system prior to hearing the voice produced by the MT-SPS system. Likely, the listeners who heard the WCS system voice first more fully appreciated a desire to have a voice sound like an individuals' own natural speech because they first heard synthetic speech from the WCS system that was generally less intelligible and less natural than the SPS system.

When asked about preferences regarding the experimental voices used in this study, participants tended to prefer the voice from the SPS system over the WCS system, which is consistent with prior reports of the perspectives of individuals who had banked their voices (e.g., 164 listeners preferred a voice from SPS, while 90 preferred WCS in Lilley et al., 2020). However, an order effect was detected for preference ratings in both the mildly and moderately dysarthric conditions: When the WCS voice was heard first, 85% of participants preferred the SPS voice, whereas when the SPS voice was heard first, preferences were almost evenly split and only 52% chose the SPS voice. Notably, the same standardized passage was used in both the

MT-SPS and MT-WCS stimuli to control for the effect of the phonetic content of the passage and presentation was counterbalanced. However, the consistent passage likely allowed the MT-SPS voice (which was rated as more intelligible and more natural than MT-WCS) to serve as a primer for the MT-WCS passage. Familiarity increases intelligibility, and listeners generally prefer passages in which they know what is being said (e.g., Borrie et al., 2012; Holmes et al., 2021; Hustad & Cahill, 2003). In this case, hearing the MT-WCS output first provided some familiarity with the passage, and the SPS output likely benefited from both that familiarity and its increased intelligibility, so most listeners preferred it. By contrast, when the MT-SPS output was heard first, it did not gain the familiarity benefit, but was more intelligible at the onset. Then when the WCS output was subsequently heard, it likely benefited from familiarity making its discontinuities less prominent to the listeners and resulting in it being preferred almost as often as the SPS output. Hence, the order effect can likely be attributed to the combined effects of familiarity and recency (Gibson et al., 1996), and also suggests that while participants had clear preferences for voices that sound more like natural speech, intelligibility remains very important.

Attitudes

Regardless of dysarthria severity, participants' attitudes were slightly more positive toward the voices produced via SPS than WCS for each of the four attitude items. Notably, this is a preliminary investigation involving only four survey items selected from an existing attitude measure and only examined four components of attitudes (i.e., comfort, pity, confidence, and willingness to interact with the AAC user). Similar to other studies of attitudes toward individuals who use AAC to communicate (e.g., Hyppa-Martin et al., 2021), reported attitudes were generally positive regardless of synthesis system or dysarthria severity (i.e., at least a rating of 3 on a 1-to-5 rating scale), with one exception: participants who listened to the mildly

dysarthric voice produced via MT-WCS indicated feeling slightly sorry for the person who communicated using that voice (mean rating 2.96). Individuals who experience disabilities have reported negative attitudes directed toward them (McCarthy & Light, 2005; Shaver et al., 1989) and selecting interventions (and voices) associated with more positive attitudes may be important to support social inclusion among individuals who use AAC to communicate.

Limitations and Future Research Directions

There are several limitations associated with this study that warrant future empirical examinations. First, this study used two specific synthesis systems, one SPS and one WCS system in current clinical use. While we believe that crucial differences between these WCS and SPS systems (i.e., more natural voice quality versus more natural speech timing and intonation) are characteristic of these classes of synthesizers, results from this study may not generalize to other SPS and WCS systems, particularly in terms of the magnitude of the effects observed. Future research must examine the relationships between the continually advancing options for speech synthesis and listener perceptions to identify affordable, accessible, high-quality voice banking options for individuals with ALS.

Our examinations of intelligibility involved semantically unpredictable sentence that are unlikely to occur in natural conversation and involved listening opportunities over a short period of time. Ultimately, intelligibility should be examined within the functional context of real-life, daily communication exchanges among a variety of actual communication partners. Intelligibility ratings increase with familiarity (Holmes et al., 2021) and future studies should examine intelligibility (and also naturalness and preferences) over time.

The majority of participants in this study were unfamiliar with AAC and, as far as we know, none were familiar with the individuals diagnosed with ALS whose synthetic speech was

used as stimuli. It is possible that listeners who are family and friends may have different perceptions about output from synthesis systems. Likely, family members may have stronger preferences for a synthetic voice to sound like the individuals' own natural voice (i.e., maintaining vocal identity) when compared to listener preferences about unfamiliar individuals. While this warrants future research, as life-extending advancements are made in treatment of ALS, less familiar people may represent an increasingly important pool of communication partners.

To our knowledge, this is the first study to employ such a large group of naïve adult listeners as participants. This large sample represents a strength of the study, but a limitation is that the demographic characteristics of the participants do not mirror the demographics of the population of the US. Participants represented a broad range of ages and educational backgrounds. However, participants who reported ethnic backgrounds such as Asian, Black/African American, Native Hawaiian/Pacific Islander, Hispanic, American Indian/Alaska Native, were minimally represented, which prevented potentially helpful analyses related to perceptions of individuals from various demographic groups. Additionally, the individuals whose speech was used to create the study stimuli were both White female monolingual English speakers. Recent and ongoing research (e.g., Westley et al., 2019) is examining voice banking among diverse populations, which is absolutely needed if interventionists are to meet the needs of an increasingly diverse clinical population.

The present examination of attitudes was preliminary and future research should continue to explore the effect of synthesis system characteristics, as well as other variables, on potential communication partners' attitudes toward individuals who use AAC to communicate. Given the long history of negative attitudes directed toward individuals with disabilities (e.g., Shaver et al.,

1989) and the need to support social inclusion and acceptance, any steps to improve attitudes toward persons experiencing severe communication disabilities are important. Future research should examine if similar results are found if a more complete survey of attitudes was used, would ideally include observations of actual behaviors (rather than reported behaviors), and could examine perceptions among medical professionals and family members who are tend to be frequent communication partners of individuals with ALS. As speech synthesis technology continues to improve, we should expect to see new, possibly more expressive synthesis supplant the type of SPS system used for the present study, which seems likely to affect listener attitudes (e.g, Aylett et al., 2017). Finally, it seems possible that as exposure to synthetic speech becomes increasingly ubiquitous (e.g., Siri), attitudes toward those who use it to communicate may improve. Explorations of these issues could serve as a valuable attitude intervention study.

This study provided insights about how one's natural speech behaviors can influence intelligibility of synthetic speech, given that the voices created from moderately dysarthric speech were more intelligible than the voices made from the mildly dysarthric speech. Future research should examine advancements to synthesis software to compensate for various aspects of dysarthria to make voice banking a viable option for all individuals with dysarthria, regardless of severity. Anecdotally, it is common for family members to ask whether a custom synthetic voice can be created from old voicemails or existing video recordings created before the onset of ALS. Most people can readily access a plethora of short recordings of an individual's speech, thanks to social media and mobile recording devices. Future research should focus on synthesis options to create custom, natural, intelligible synthetic voices from these existing recordings.

Finally, voice banking options to provide custom, personalized synthetic voices should be developed for and offered to individuals with conditions other than dysarthria secondary to ALS,

including persons facing laryngectomy, individuals with developmental disabilities who did not develop intelligible speech to record and bank (e.g., cerebral palsy, autism), and individuals who experience co-occurring conditions such as apraxia (which may complicate the individual's ability to produce the target utterances necessary for voice banking). Individuals with complex communication needs who wish to communicate using a custom synthetic voice that sounds recognizably like their own natural speech and is reflective of their age, social, linguistic, cultural, and geographic influences would benefit from continued investigation of these issues.

Conclusion

Understandably, a majority of funding that supports ALS research is directed at identifying a cure (ALS Association, 2021). Until a cure is developed, ensuring that individuals diagnosed with ALS can communicate effectively and with a voice that maintains their identity for the duration of their lives represents an important goal of AAC and voice banking. Notably, this study provides an example of the role that interdisciplinary collaborations play in achieving that goal which, in this case, required scientists with linguistics, psychology, and computational skills to collaborate with practicing clinician-researchers. Interprofessional practices focused on the development of AAC technologies, including custom synthetic voices, are crucial in moving toward a more inclusive society in which individuals facing severe communication disabilities are active participants.

References

Acapela Group. (2022). *My-own-voice*. https://mov.acapela-group.com/

ALS Association. (2020). *FYI: A guide to voice banking services*. https://www.als.org/navigating-als/resources/fyi-guide-voice-banking-services

ALS Association. (2021). *A look back at over $16 million in research grants awarded during 2018*. http://web.alsa.org/site/PageNavigator/blog_050319.html

ALS Association. (2022). *Understanding ALS*. https://www.als.org/understanding-als

Aylett, M, Vinciarelli, A., & Wester, M. (2017). Speech synthesis for the generation of artificial personality, *IEEE Transactions on Affective Computing*, *11*(2), 361-372. https://doi.org/10.1109/TAFFC.2017.2763134

Baird, A., Parada-Cabaleiro, E., Hantke, S., Burkhardt, F., Cummins, N., &. Schuller, B. (2018). The perception and analysis of the likeability and human likeness of synthesized speech. *Proceedings of INTERSPEECH*, 2863–2867. https://doi.org/10.21437/Interspeech.2018-1093

Ball, L., Beukelman, D., & Pattee, G. (2004). Acceptance of augmentative and alternative communication technology by persons with amyotrophic lateral sclerosis. *Augmentative and Alternative Communication*, *20*(2), 113–122. https://doi.org/10.1080/0743461042000216596

Benoit, C., Grice, M., & Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. *Speech Communication*, *18*(4), 381–392. https://doi.org/10.1016/0167-6393(96)00026-X

Black, A., Zen, H., & Tokuda, K. (2007). *Statistical Parametric Speech Synthesis*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.154.9874&rep=rep1&type=pdf

Borrie, S., McAuliffe, M., Liss, J., Kirk, C., O'Beirne, G., & Anderson, T. (2012). Familiarisation conditions and the mechanisms that underlie improved recognition of dysarthric speech. *Language and Cognitive Processes*, *27*(7–8), 1039–1055. https://doi.org/10.1080/01690965.2011.610596

Bunnell, H. T. (2010). Crafting small databases for unit selection TTS: Effects on intelligibility. *Proceedings of the 7th ISCA Speech Synthesis Workshop*, 40–44. https://www.isca-speech.org/archive_v0/ssw7/papers/ssw7_040.pdf

Bunnell, H. T., & Lilley, J. (2007, August 22-24). *Analysis methods for assessing TTS intelligibility* [Paper presentation]. 6th ISCA Workshop on Speech Synthesis, Bonn, Germany. https://www.isca-speech.org/archive_open/archive_papers/ssw6/ssw6_374.pdf

Bunnell, H. T., & Pennington, C. (2010). Advances in computer speech synthesis and implications for assistive technology (pp. 71–91). In J. Mullennix & S. Stern (Eds.), *Computer synthesized speech technologies: Tools for aiding impairment*. IGI Global. https://doi.org/10.4018/978-1-61520-725-1

CereProc Ltd. (2022). *CereProc Text-to-Speech*. https://www.cereproc.com/en/cerevoice-me

Capes, T., Coles, P., Conkie, A., Golipour, L., Hadjitarkhani, A., Hu, Q., Huddleston, N., Hunt, M., Li, J., Neeracher, M., Prahallad, K., Raitio, T., Rasipuram, R., Townsend, G., Williamson, B., Winarsky, D., Wu, Z., & Zhang, H. (2017). Siri on-device deep learning-guided unit selection text-to-speech system. *Proceedings of INTERSPEECH*, 4011–4015. https://doi.org/10.21437/INTERSPEECH.2017-1798

Costello, J. (2018). Last Words, last Connections: How AAC can support children facing end of life. *ASHA Leader*. https://doi.org/10.1044/leader.FTR2.14162009.8

Creer, S., Cunningham, S., Green, P., & Yamagishi, J. (2013). Building personalised synthetic voices for individuals with severe speech impairment. *Computer Speech and Language*, *27*(6), 1178–1193. https://doi.org/10.1016/j.csl.2012.10.001

Doyle, M., & Phillips, B. (2001). Trends in augmentative and alternative communication use by individuals with amyotrophic lateral sclerosis. *Augmentative and Alternative Communication*, *17*(3), 167–178. https://doi.org/10.1080/aac.17.3.167.178

Eagly, A., & Chaiken, S. (2007). The advantages of an inclusive definition of attitude. *Social Cognition*, *25*(5), 582–602. https://doi.org/10.1521/soco.2007.25.5.582

Fairbanks, G. (1941). Voice and articulation drillbook. *The Laryngoscope, 51*(12), 1141. https://doi.org/10.1288/00005537-194112000-00007

Gibson, E., Pearlmutter, N., Canseco-Gonzalez, E., & Hickok, G. (1996). Recency preference in the human sentence processing mechanism. *Cognition*, *59*(1), 23–59. https://doi.org/10.1016/0010-0277(95)00687-7

Gorenflo, C., & Gorenflo, D. (1991). The effects of information and augmentative communication technique on attitudes toward nonspeaking individuals. *Journal of Speech and Hearing Research*, *34*(1), 19–26. https://doi.org/10.1044/jshr.3401.19

Hanson, E., Yorkston, K., & Britton, D. (2011). Dysarthria in amyotrophic lateral sclerosis: A systematic review of characteristics, speech treatment, and augmentative and alternative communication options. *Journal of Medical Speech-Language Pathology*, *19*(3), 12–31.

Hinterleitner, F. (2017). *Quality of synthetic speech: Perceptual dimensions, influencing factors, and instrumental assessment*. Springer.

Holmes, E., To, G., & Johnsrude, I. (2021). How long does it take for a voice to become familiar? *Psychological Science*, *32*(6), 903–915. https://doi.org/10.1177/0956797621991137

Honorof, D., McCullough, J., & Somerville, B. (2000). *Comma gets a cure.* Newcastle University. https://research.ncl.ac.uk/necte2/documents/comma.pdf

Hunt, A., & Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. https://doi.org/10.1109/ICASSP.1996.541110

Hustad, K., & Cahill, M. (2003). Effects of Presentation Mode and Repeated Familiarization on Intelligibility of Dysarthric Speech. *American Journal of Speech-Language Pathology*, *12*(2), 198–208. https://doi.org/10.1044/1058-0360(2003/066)

Hyppa-Martin, J., Chen, M., Janka, E., & Halverson, N. (2021). Effect of partner reauditorization on young adults' attitudes toward a child who communicated using nonelectronic augmentative and alternative communication. *Augmentative and Alternative Communication*, *37*(2), 141–153. https://doi.org/10.1080/07434618.2021.1916075

Hyppa-Martin, J., Friese, J., & Barnes, C. (2017). *Voice banking for individuals with ALS: Tips and pointers for successful, low-cost voice banking.* [Poster presentation]. American Speech-Language-Hearing Association Convention, Los Angeles, CA, United States.

IEEE Subcommittee. (1969). *Harvard Sentences.* Columbia University. https://www.cs.columbia.edu/~hgs/audio/harvard.html

Kraus, S. (1995). Attitudes and the prediction of behavior: A metaanalysis of the empirical literature. *Personality & Social Psychology Bulletin*, *21*(1), 58–75. https://doi.org/10.1177/0146167295211007

Kraker, D. (2018). *ALS robbing them of speech, but they won't be silenced*. MPR News. https://www.mprnews.org/story/2018/08/06/voice-banking-preserves-voices-of-people-who-might-lose-ability-to-speak-als

Kühnlein, P., Gdynia, H., Sperfeld, A., Lindner-Pfleghar, B., Ludolph, A., Prosiegel, M., & Riecker, A. (2008). Diagnosis and treatment of bulbar symptoms in amyotrophic lateral sclerosis. *Nature Clinical Practice Neurology*, *4*(7), 366–374. https://doi.org/10.1038/ncpneuro0853

Kuligowska, K., Kisielewicz, P., & Włodarz, A. (2018). Speech synthesis systems: Disadvantages and limitations. *International Journal of Engineering & Technology*, *7*(2.28), 234. https://doi.org/10.14419/ijet.v7i2.28.12933

Lilley, J., Hyppa-Martin, J., & Bunnell, H. T. (2020). A large-scale comparison of the intelligibility of unit-selection and deep-neural-network parametric synthetic voices generated from dysarthric speech. *The Journal of the Acoustical Society of America*, *148*(4), 2582–2582. https://doi.org/10.1121/1.5147169

Ling, Z., Wu, Y., Wang, Y., Qin, L., & Wang, R. (2006). *USTC System for Blizzard Challenge 2006 an Improved HMM-based Speech Synthesis Method*. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.130.7143&rep=rep1&type=pdf

Mayo, C., Clark, R., & King, S. (2011). Listeners' weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis. *Speech Communication*, *53*(3), 311–326. https://doi.org/10.1016/j.specom.2010.10.003

McCarthy, J., & Light, J. (2005). Attitudes toward individuals who use augmentative and alternative communication: Research review. *Augmentative and Alternative Communication*, *21*(1), 41–55. https://doi.org/10.1080/07434610410001699753

Mills, T., Bunnell, H. T., & Patel, R. (2014). Towards personalized speech synthesis for augmentative and alternative communication. *Augmentative and Alternative Communication*, *30*(3), 226–236. https://doi.org/10.3109/07434618.2014.924026

Morise, M., Yokomori, F., & Ozawa, K. (2016). WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems, E99-D*(7), 1877–1884.

Mulligan, M., Carpenter, J., Riddel, J., Delaney, M. K., Badger, G., Krusinski, P., & Tandan, R. (1994). Intelligibility and the acoustic characteristics of speech in amyotrophic lateral sclerosis (ALS). *Journal of Speech, Language, and Hearing Research*, *37*(3), 496–503. https://doi.org/10.1044/jshr.3703.496

Nusbaum, H., Francis, A., & Henly, A. (1995). Measuring the naturalness of synthetic speech. *International Journal of Speech Technology*, *1*(1), 7–19. https://doi.org/10.1007/BF02277176

Nemours Children's Health. (2021). *ModelTalker: Creating Personal Voices For All*. https://www.modeltalker.org/

Shaver, J., Curtis, C., & Strong, C. (1989). The modification of attitudes toward persons with disabilities: Is there a best way? *International Journal of Special Education,* 4, 33–57.

Shen, J., Pang, R., Weiss, R., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). *Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions*. IEEE

International Conference on Acoustics, Speech and Signal Processing. https://doi.org/10.48550/arXiv.1712.05884

Tomik, B., Krupinski, J., Glodzik-Sobanska, L., Bala-Slodowska, M., Wszolek, W., Kusiak, M., & Lechwacka, A. (1999). Acoustic analysis of dysarthria profile in ALS patients. *Journal of the Neurological Sciences*, *169*(1–2), 35–42. https://doi.org/10.1016/S0022-510X(99)00213-0

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). *WaveNet: A generative model for raw audio.* http://arxiv.org/abs/1609.03499

Veaux, C., Yamagishi, J., & King, S. (2011). Voice banking and voice reconstruction for MND patients. *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility.* https://doi.org/10.1145/2049536.2049619

Westley, M., Sutherland, D., & Bunnell, H. T. (2019). Voice banking to support people who use speech-generating devices: New Zealand voice donors' perspectives. *Perspectives of the ASHA Special Interest Groups*, *4*(4), 593–600. https://doi.org/10.1044/2019_PERS-SIG2-2018-0011

Wu, Z., Watts, O. & King, S. (2009). Merlin: An open source neural network speech synthesis system. *Proceedings of the 9th ISCA Speech Synthesis Workshop,* 218–233. http://ssw9.talp.cat/papers/ssw9_PS2-13_Wu.pdf

Yamagishi, J., Veaux, C., King, S., & Renals, S. (2012). Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology*, *33*(1), 1–5. https://doi.org/10.1250/ast.33.1

Zen, H., Tokuda, K., & Black, A. (2009). Statistical parametric speech synthesis. *Speech Communication*, *51*(11), 1039–1064. https://doi.org/10.1016/j.specom.2009.04.004

Zhang, H., Chen, L., Tian, J., & Fan, D. (2021). Disease duration of progression is helpful in identifying isolated bulbar palsy of amyotrophic lateral sclerosis. *BMC Neurology, 21*(1), 405–405. https://doi.org/10.1186/s12883-021-02438-8