June 2017

Perceptual Differences in Natural Speech and Personalized Synthetic Speech

Katherine Overton
University of South Florida, overtonk@mail.usf.edu

Perceptual Differences between Natural Speech and Personalized Synthetic Speech

by

Katherine Overton

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science
Department of Communication Sciences and Disorders
College of Behavioral and Community Sciences
University of South Florida

Major Professor: Michelle Bourgeois, Ph.D., CCC-SLP
R. Michael Barker, Ph.D.
Kyna Betancourt, Ph.D., CCC-SLP

Date of Approval:
June 21, 2017

Keywords: ModelTalker, message banking, amyotrophic lateral sclerosis (ALS), augmentative alternative communication (AAC)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The purpose of this study was to determine what perceptual differences existed between a natural recorded human voice and a synthetic voice that was created to sound like the same voice. This process was meant to mimic the differences between a voice that would be used for Message Banking and a voice that would be created by the ModelTalker system. Forty speech pathology graduate students (mean age = 23 years) rated voices on clarity, naturalness, pleasantness, and overall similarity. Analysis of data showed that the natural human voice was consistently rated as more natural, clear, and pleasant. In addition, participants generally rated the two voices as very different. This demonstrates that, at least in terms of perception, using the method of Message Banking results in a voice that is overall perceived more positively than the voice created using ModelTalker.

INTRODUCTION

Amyotrophic lateral sclerosis (ALS) is a progressive neurological disease that typically results in death within 2 to 5 years (Miller et al., 2009). There is no known cure for ALS, and treatment typically focuses on symptom management (Miller at al., 2009). Speech language pathologists can become part of the ALS care team in order to address both speech and swallowing impairments. Communication, in particular, can be impaired by the presence of dysarthria (Miller et al., 2009). There are two main types of ALS: bulbar and spinal. Bulbar ALS develops in the brainstem and is characterized by difficulties in speech, swallowing, and breathing in the early stages. Spinal ALS develops in the spinal cord and typically begins in the patient's limbs. Speech symptoms in bulbar onset ALS commonly manifest as a decreased speaking rate and a progression from dysarthria to complete anarthria (Ball, Beukelman & Mirenda, 2005). Ball and colleagues (2005) outlined a staging hierarchy for the progression of speech impairment: (1) no detectible speech impairment, (2) obvious impairment but intelligible, (3) decreased intelligibility, (4) residual speech, and (5) loss of functional speech.

According to estimates from multiple studies, between 75% and 95% of people with ALS cannot communicate without assistance at the time of their death (Ball et al., 2005). Almost all people with ALS present with some level of dysarthria during the course of their disease (Ball et al., 2005). It is recommended that augmentative-alternative communication (AAC) referrals should begin when the client has an average speaking rate of between 100 and 120 words per minute, which indicates a speech intelligibility of below 90% or when

communication suffers in "adverse listening situations" (Ball et al., 2005). Furthermore, the staging of AAC intervention can correspond with the staging hierarchy of speech impairment progression previously outlined (Ball, et al., 2005).

Throughout all five stages, clients should be encouraged to utilize both high and low tech AAC, especially since low tech AAC use can increase towards the end of life (Ball et al., 2007). During Stage One, intervention focuses on providing general information about services that may be needed in the future and consistently monitoring speech rate and intelligibility to identify progression to Stage Two as accurately as possible. The length of Stage One can depend on the type of ALS onset, with longer duration for spinal onset and shorter duration for bulbar onset. No AAC implementation is required at this stage. Intervention during Stage Two focuses on development of strategies for both the individual and their communication partners to maximize communication. Stage Two may be when the client reaches the threshold for initiating AAC assessment, but the client will still require no AAC at this time.

Stage Three involves increasingly more complex strategies including monitoring of speech rate and improving breath control. An AAC assessment should be completed by this phase in order to provide additional and more specific information about their communication options and the use of AAC during the progression of ALS. Individuals will typically begin using a combination of high and low tech AAC during this stage. For example, high tech options such as a speech generating device might be helpful during telephone conversations, and low tech options, such as a communication board with pictures could facilitate conversations with familiar partners.

The fourth stage begins when AAC becomes central to communication rather than merely a supplement to natural speech; however, natural speech can still be used for at least a portion of communication. At this point, intervention can focus on education about how to continue to maximize natural speech, while also implementing AAC. Electronic methods of communication, such as speech generating devices, may become very important, especially in difficult communication situations, such as speaking in noisy environments and talking on the telephone.

Finally, at stage five, communication using natural speech is no longer feasible, and AAC becomes the sole mode of communication. The development of efficient low tech strategies, such as the use of a yes/no system to communicate basic needs, becomes very important, while high tech strategies, such as eye-gaze and speech generating devices, can be implemented for more complex communication. This stage is the point at which mechanical ventilation may be required and motor function is severely impaired, leading to either modification of the pre-existing AAC system or even a completely new device (Ball et al., 2007).

AAC needs also can vary based on other aspects of disease progression, such as decline in motor skills. This distinction can be especially relevant when determining the AAC needs of an individual with spinal vs. bulbar disease onset. Typically, people with bulbar onset ALS can use direct selection for significantly longer periods of time than those with spinal ALS due to preserved motor control of the limbs (Ball et al., 2005). Individuals with spinal ALS are more likely to require alternative methods of selection, such as eye gaze, earlier due to limb involvement (Ball et al., 2005).

As ALS progresses, alternative augmentative communication (AAC) eventually becomes the patient's primary mode of communication. Thus, AAC is an integral part of the life of a person with ALS. Ball, Beukelman, and Pattee (2004) investigated the acceptance of AAC technology in persons with ALS. They found that 90% of the participants immediately accepted AAC technology, 6% accepted it after some delay, and only 4% refused to use AAC technology. Refusal of AAC technology was attributed to cognitive impairments caused by ALS-related dementia, which results in behavioral and cognitive changes that may impair the ability to fully understand the changes in their communication needs and the implications of their disease. Participants who did not accept AAC refused all levels of AAC including low tech and no technology devices and strategies. Delayed acceptance, on the other hand, was caused by a variety of factors, including resistance by family, physician, or the person with ALS. Family resistance was characterized by the belief that AAC was unnecessary and possibly even a condemnation of their abilities as caregivers. Physician resistance occurred in one instance in which the physician counseled the family to accept the communication impairments as they were, rather than attempting to ameliorate them using technology. Finally, resistance from the individual with ALS was commonly due to denial of the illness or denial of the need for AAC due to their perception that their speech intelligibility was adequate. However, all three of these types of concerns were eventually overcome and AAC devices and strategies were implemented. This overwhelming acceptance of AAC demonstrates the pivotal role that AAC plays in the lives of people with ALS. The desire to communicate can frequently outweigh any other concerns that may exist.

In addition, the time period in which people with ALS use their devices is telling. Ball et al. (2007) examined the duration of AAC use in people with ALS and compared it to factors, such as age, gender, type of ALS (spinal or bulbar), use of mechanical ventilation and feeding tubes, and timeliness of referral. They found that age, gender, and type of ALS had no statistically significant effect on duration. However, they found that timeliness of referral and the use of life extending technologies (mechanical ventilation and feeding tubes) did affect the duration of use. The mean duration of use across all participants was 28.4 months, with higher durations for those using mechanical ventilation and/or PEG tubes shorter durations for those who did not receive a timely referral for AAC evaluation. They also found that all participants used some form of AAC technology within days of their death (Ball et al. 2007). These findings further highlight the importance of AAC in the lives of people with ALS.

Several other barriers to the acceptance and use of AAC include the quality of the voices used in the speech generating devices (SGD) and the time and effort involved in creating an acceptable synthetic voice for use with an SGD. Speech generating devices can come in a variety of forms, from iPad apps to entire devices dedicated solely to communication. All of these come with their own standard synthesized voices, sometimes with several options based on age and gender. Unfortunately, with many people using similar devices and limited options of voices, many people use the same voice, which can cause concerns about maintaining one's individuality and identity while using an SGD. Based on these concerns, new technologies have expanded the options of individuals who require SGDs. For example, a program called ModelTalker uses concatenative speech synthesis to create a personalized voice from a corpus of 1600 sentences recorded by the person with ALS (Mills,

Bunnell & Patel, 2014). Ideally, this program would provide a feasible alternative to professional voices that would maintain personal vocal characteristics but also have the ability to communicate an unlimited number of messages. However, as noted above, the process requires the recording of 1600 sentences, which is a time consuming undertaking. The time constraints and limited energy that a person with ALS lives with could make this all but impossible. Thus, it becomes important to evaluate whether the synthetic voices created are worth the time and effort. An additional concern with ModelTalker can be the timeliness of referral, as noted in Ball et al. 2007. If disease progression has passed into either stage two or three, speech quality may no longer be conducive to creating a synthesized voice.

The intense and complex process of synthetic speech creation is a major barrier to the creation of quality synthetic voices based on a corpus of non-professional recordings. Andersson, Yamagishi, and Clark compared synthetic voices created through the commonly used method of eliciting samples from sentences being read aloud, and synthetic voices created using conversational samples (2011). They found that the conversational samples created voices that were perceived as more conversational and more natural. The conversational samples were created from over seven hours of recording of a conversation with a professional voice actor done in a professional recording studio. These circumstances, while producing better quality voices, are not feasible outside of a closely supervised professional setting, meaning that this technique, as of yet, cannot be implemented in a program like ModelTalker or Acapela "my-own-voice." Ilves and Surakka (2013) compared how listeners perceived emotion in three different types of synthetic voices: formant synthesis, unit selection synthesis, and diphone synthesis. Even the voice perceived as the most mechanical

(formant synthesis) was still perceived as conveying some level of emotion, but the most natural voice was also perceived as conveying the most emotion. Unit selection was perceived as more natural than diphone synthesis; however, unit synthesis requires a significantly larger corpus of recordings and computing power to create. As with conversational sampling vs. reading aloud, the ideal method of creating a synthetic voice professionally cannot be replicated under the circumstances necessary to create personalized voices. The ModelTalker program uses an adapted form of diphone synthesis to create its voices (Mills, Bunnell & Patel 2014).

An alternate option for people who may not be an appropriate candidate for voice banking is message banking. For message banking, participants can record a series of commonly used and/or personally important messages, which can be later integrated into a speech generating device. While this method results in a limited number of output possibilities, the personalization options (sports cheers, profanities, prayers) can at least partially ameliorate that concern. In addition, recorded natural speech is obviously closer to actual natural speech than a synthesized version of natural speech. Stern, Mullenix, and Wilson found that while an understanding of disability may improve the perception of synthetic voices, people still rate natural human voices significantly higher than synthetic voices in terms of both favorability and persuasiveness (2002).

In order to empirically assess the relative merits of the two options, it becomes important to assess the quality of the voices produced. Many studies have compared natural voices to synthesized voices, but as of yet, this methodology has not been applied to the creation of personalized synthetic voices. A frequently used method is qualitative comparison

of listener perception (Stern, Mullenix & Wilson, 2002; Mullenix, Stern, Wilson & Dyson, 2003; Cote-Giroux et al., 2011). All three studies used semantic differential scales to measure comparative perception of multiple parameters. Cote-Giroux et al. compared synthetic voices on four parameters: warm vs. cold, soft vs. hard, monotone vs. expressive, and smooth vs. bumpy (2011). For perception of speech quality, Stern, Mullenix, and Wilson measured seven characteristics: "loud voice–soft-spoken voice, deep voiced–squeaky voiced, fast speaking–slow speaking, heavy accent–faint accent, talked too long–didn't talk long enough, heavy nasality–faint nasality, and monotone–lively" (2002). Mullenix, Stern, Wilson, and Dyson used the same scale (2003). Andersson, Yamagishi, and Clark evaluated synthetic voices based on listener perception of naturalness and conversational speaking style (2011). Richter, Ball, Beukelman, Ullman, and Lasker measured listener perception of modes of storytelling by measuring perception of communicator competence, effectiveness of communication, listener comfort, willingness to participate in future conversation, and comprehension of message (2009). Ilves and Surakka has listeners assess pleasantness, naturalness, and clarity (2013).

Each study outlined above used these perceptual measures to study a different aspect of synthetic speech. The study by Stern, Mullenix, and Wilson (2002) and the study by Mullenix, Stern, Wilson and Dyson (2003) both looked at the persuasiveness of synthetic voices. The former focused the relationship between the perception of disability and persuasiveness, and the latter investigated the effect of gender on the persuasiveness of synthetic voices. Stern, Mullenix, and Wilson found that while natural speech was rated significantly higher than synthetic speech, the ratings of synthetic speech increased when the speaker was perceived as disabled (2002). Mullenix, Stern, Wilson, and Dyson also provided

evidence that natural speech is consistently perceived as more persuasive than synthetic speech (2003).

While Mullenix, Stern, Wilson, and Dyson found some differences in perception of voices of different gender, Cote-Giroux et al. found no effect on perception due to that factor (2003; 2011). Cote-Giroux et al. compared one male human voice and nine synthetic voices of both genders on both perceptual characteristics and intelligibility (2011). Two of the synthetic voices were found to be statistically as intelligible as the human voice in all conditions; three additional voices were as intelligible as the human voices when words were presented in context. One of the synthetic voices that was found to be as intelligible as the human voice was also given a statistically similar positive perceptual score to the human voice. These findings indicate that the quality of professionally produced synthetic voices are coming close to approximating natural human speech in both intelligibility and perceptual quality.

Many factors can affect the perception of synthetic speech. Delogu, Conte, and Sementina found that listener ratings of synthetic speech improved with reported exposure to synthetic voices (1998). The ability to speak in a conversational style can be significantly correlated with the perception of naturalness (Andersson, Yamagishi & Clark, 2012). Ilves and Surakka found that while even mechanical synthetic voices can express emotion, the perception of emotion can increase with the naturalness of the voice (2012). All of these factors make the creation of high quality synthetic voices difficult, but some factors such as the increased exposure to synthetic voices are already happening naturally as overall technology improves.

The only study found that directly addressed the perception of synthetic speech by people with ALS was completed by Richter, Ball, Beukelman, Lasker, and Ullman (2009). They used two separate studies to investigate perceptions of storytelling for people with ALS. One of the studies looked at the method by which the story was relayed, including no AAC (natural speech), low tech AAC (communication book), and high tech AAC (synthetic speech). In both studies, perception was based on people with ALS, caregivers, and naïve listeners. The results of the studies indicated that while there was strong agreement between listener groups, and overall, listeners significantly preferred the use of AAC to natural but unintelligible speech. As with the studies pertaining to ALS that were previously discussed, it demonstrates the importance of appropriate AAC to people with ALS, and also, that synthetic speech is a viable and important option. In addition, while the acoustic quality of personalized synthetic voices has been evaluated, there has not yet been research in quality related to listener perception of voices. By comparing the synthetic voice to the voice from which is was created, the quality of the end product can be better evaluated. As discussed previously, creating a personalized voice takes a significant commitment of both time and effort, so a better understanding of the outcome of that time and effort will improve the ability to make informed decisions about whether to pursue this method.

The current study aimed to evaluate the quality of the personalized synthetic voice in comparison to the individual's recorded natural voice. Quality was assessed through perceptual characteristics, such as prosody and intelligibility, as rated by unfamiliar listeners. Although quality of synthetic voices has been studied, little research exists on listener perception of these personalized voices. The research questions were:

1. What, if any, difference will there be in intelligibility between the voice that would be used for Message Banking and the voice that is produced by ModelTalker?
2. What are the differences in perception as measured by the perceptual characteristics (pleasantness, clarity and naturalness) of two different types of message types?
3. Overall, how similar or different will the voices be perceived to be?

It was hypothesized that ModelTalker samples will be less intelligible and will be rated to be less clear, pleasant, and natural than the Message Banking samples. It was also hypothesized that the two message types will be perceived to be different.

METHODS

Participants

Forty graduate students in their first year of the Speech Language Pathology program at the University of South Florida participated in the study. There were 39 female participants and one male participant between the ages of 21 and 29 years (mean = 23.2; s.d. = 2.07). Participants were recruited from two classes. Inclusion criteria included enrollment in the master's degree program in Speech Pathology. Exclusion criteria included a severe hearing loss based on self-report. One participant was excluded due to self-report. IRB consent was waived in order to maintain anonymity of participants.

Familiarity with Synthetic Voices

Participants also self-rated themselves on their perceived familiarity with synthetic voices. Of the forty participants, twenty-one (52.5%) rated themselves as "casual," seventeen (42.5%) rated themselves as "familiar," and two (5%) participants rated themselves as "knowledgeable." None of the participants rated themselves as "unfamiliar" or "expert." Participants were also asked about the source of their familiarity. Most participants identified cellphone (n=38), GPS (n=37), and computer (n=31) as sources of exposure to synthetic voices. Five participants added AAC and clinic as sources of exposure, and three participants listed research.

Materials

Voices

Participants listened to two different types of recorded voices. Voice A was a recording of the natural voice of a 22 year-old female student meant to represent Message Banking. Voice B was a ModelTalker voice created by the same female. Both the ModelTalker voice and the recordings of Voice A were recorded using a Sennheiser PC 36 headset with a built-in USB microphone. Voice A was recorded using the Pratt program, and Voice B was created using recordings on the ModelTalker website.

Stimulus Recordings

There were a total of ten stimulus phrases each with five syllables. In order to control for frequency and other acoustic variables, all recordings were analyzed acoustically using Praat (Boersma & Weenink, 2015). Any significant disparities found in fundamental frequency (F0), intensity, or rate were edited using built-in settings on ModelTalker. In order to maintain consistency of sample and decrease interference, all adjustments were made using the Synthesis Parameters function in the ModelTalker desktop speech synthesizer, which would be accessible to any client or clinician. Intensity measures were matched within ±3 dB by decreasing the "Speaking Volume" parameter from 100 to 10. Speaking rate was matched by measuring words per minute (WPM) in the Grandfather Passage. To match the rate of the original speaker, the "Speaking Rate" parameter was decreased from 120 to 10.

Table 1. Phrases Recorded for Audio Sample

I love you so much.	Can you help me please?
How are you doing?	I like to read books.
I am not done yet.	

Measures

Demographic Questionnaire. Each participant was given a brief demographic questionnaire including age, gender, and ethnicity (see Appendix A for an example of the demographic form).

Rating Forms. Each participant was given a packet of 10 rating forms; one for each sentence to be rated. Each page consisted of numbered items to first write down the sentence that was heard, then to use the semantic differential ratings line to mark their perception of each of 3 perceptual characteristics (see Appendix B for an example of the Rating form). The last page of the packet displayed a semantic differential rating line for participants to rate how similar or different they perceive the two voices to be. See Appendix C for an example of the Rating form.

Design

This study compared perceptual ratings of two different voices using a within subjects comparison design. Each participant's transcriptions and ratings of both the natural voice and the synthesized version of the voice when listening to 10 audio samples (ten samples in each of two voices) presented in a randomized order were measured and tabulated.

Independent Variable

The independent variable was the type of voice used to create the stimuli, either the natural voice (Message Banking) or the synthesized version of that voice (Model Talker). Each phrase was recorded in both voices. The 10 sentences were saved into a master file in a random

order. In addition, to compare the similarity or difference between the two voices, recordings of the Grandfather passage (Darley, Aronson, & Brown, 1975) were created using both voices.

Dependent Variable

The dependent variables were 1) intelligibility, based on the accuracy of participants' transcription of each phrase; 2) the perceptual values (clarity, pleasantness, and naturalness) that were rated on a 100 mm visual analog scale (See Appendix B for an example of the rating scale); and 3) the similarity measure based on participant ratings of the similarity between the two voice conditions (See Appendix C for an example of the Similarity Rating Scale).

Procedure

The study was conducted in two classes during the 75-minute class period for first year students, with permission of the instructors. The researcher presented a brief description of the study (see Appendix E for a script of this introduction), then distributed the packets containing the Consent Form, the Demographic form, and the Rating forms. She then explained the consent forms, asking students who wished to participate to sign the forms and thanking the students who did not wish to participate as they left the room. Participants asked questions for clarification, if necessary. When there were no further questions, the experimenter presented the recorded phrases. The participants were given 30 seconds between each phrase in order to complete the necessary portions of the orthographic transcription and ratings. After all ten recordings were presented the experimenter then presented the Grandfather passage in both Voice A and Voice B. Participants were asked to rate the similarity of the two voices and provide any relevant comments. Afterwards, the experimenter answered any questions the participants had about the study.

Data Analysis

Intelligibility

Data sheets were scored for accuracy of transcription, tabulated and entered into an Excel spreadsheet. Descriptive statistics (mean, s.d.) were calculated for each individual and for each group. Percent accuracy of the transcribed sentences was compared within each subject to look for any significant differences between voice conditions using a paired sample t-test.

Perceptual Characteristics

Data from the visual analog scale was scored by measuring the participants' mark on the 100 mm line. The line was measured using a ruler from one endpoint to the mark, and this measurement was the value entered in the spreadsheet to denote the participant's rating. Based on these numbers, descriptive statistics, such as mean and standard deviation, were calculated for each characteristic. For each characteristic and voice type, the participant's mean rating was calculated. These means were compared using a paired t-test to analyze the data for significant difference. For example, a mean was calculated for Participant 1's ratings of the clarity of all audio samples of the natural voice, and that mean was compared to the mean of Participant 1's ratings of clarity for the paired audio samples of the synthetic voice. All participants mean ratings were compiled for each vocal quality (clarity, pleasantness, and naturalness) to calculate both descriptive statistics and complete a paired t-test for statistical significance. Effect size was calculated using Cohen's *d* (Cohen, 1988).

Overall Similarity

The Similarity rating data was scored by measuring the line of the visual analog scale, as described above. Ratings were entered into the spreadsheet, and descriptive statistics were calculated to measure overall rating of similarity across all participants.

Comments

All comments were compiled and then analyzed for consistent themes based on verbal prompts for comments (see Appendix E for script). Comments were gathered in a comparatively informal manner and should thus be interpreted with care.

RESULTS

Question One: Intelligibility

In order to answer the question of what, if any, difference would there be in intelligibility between the Message Banking voice and the ModelTalker voice, participants wrote the sentences they heard for each voice stimulus. Analysis of the transcription data revealed 100% accuracy of each stimulus sentence in both voice conditions.

Question Two: Perceptual Characteristics

As shown in Table 1, differences in perception as measured by the perceptual characteristics of two different message types was analyzed by comparing the participants' average ratings of the five samples of Voice A and the participants' average rating of the five samples of Voice B. Voice A was, on average, rated higher across all three characteristics. Results revealed that clarity ratings of Voice A ($M = 93.60$; $SD = 6.68$) and Voice B ($M = 85.56$; $SD = 11.06$) were significantly different ($p < 0.001$) with a large effect size ($d = 1.15$). Differences in pleasantness ratings for Voice A ($M = 90.93$; $SD = 8.69$) and Voice B ($M = 73.09$; $SD = 15.04$) were significantly different ($p < 0.001$) with a large effect size ($d = 2.12$). Finally, results revealed significant differences in naturalness ratings for Voice A ($M = 86.49$; $SD = 11.97$) and Voice B ($M = 58.32$; $SD = 20.67$) ($p < 0.001$) with a large effect size ($d = 2.38$).

Table 2. Analysis of Measures of Vocal Quality

Parameter	Mean Difference	Standard Deviation	t-value	Significance	Cohen's *d*
Naturalness	28.18	14.36	12.41	< 0.000	2.376
Pleasantness	17.84	10.29	10.96	< 0.000	2.12
Clarity	8.04	8.08	6.93	< 0.000	1.15

Reliability

A trained investigator remeasured 20% sample of the participant visual analog scale protocols to confirm inter-judge reliability. A total of 240 visual analog scales were measured. Agreement was calculated using Cohen's kappa, which resulted in near perfect agreement (k = 0.960).

Qualitative Data

Participants' qualitative comments about each stimulus sentence were compiled and analyzed. Out of 40 participants, 29 participants produced written responses in the comments section that focused specifically on the vocal quality as a notable difference between the two voices after listening to all the audio samples. Each of these written responses were coded by which voice was mentioned and what descriptor was used. Most of the written responses could be sorted under more than one code. For example: Table 2 shows the distribution of the descriptions between the two voices. There were clear differences between the comments for the voices. The Message Banking voice was most frequently described as natural, clear, and pleasant, in contrast to the ModelTalker voice that was described as robotic, unnatural and

unclear. Other descriptions of Voice B included "strained," "weak," "choppy," and "rough." Multiple participants mentioned prosody as a notable difference between the two voices.

Table 3. Most Common Descriptions by Voice

Voice A (Message Banking)	Frequency	Voice B (ModelTalker)	Frequency
Natural	9	Robotic	9
Good Prosody	5	Unnatural	7
Clear	3	Bad Prosody	6
Pleasant	2	Unclear	5

Questions Three: Overall Similarity

To determine the degree to which participants would find the two voices to be different the participants' ratings of The Grandfather Passage were compiled and averaged. As shown in Figure 1, the distribution of participants' ratings of similarity/difference ranged from 0 to 78, with 0 as different and 100 as the same. Overall across participants, the two voices were found to be different (M = 27.13, SD = 17.06).

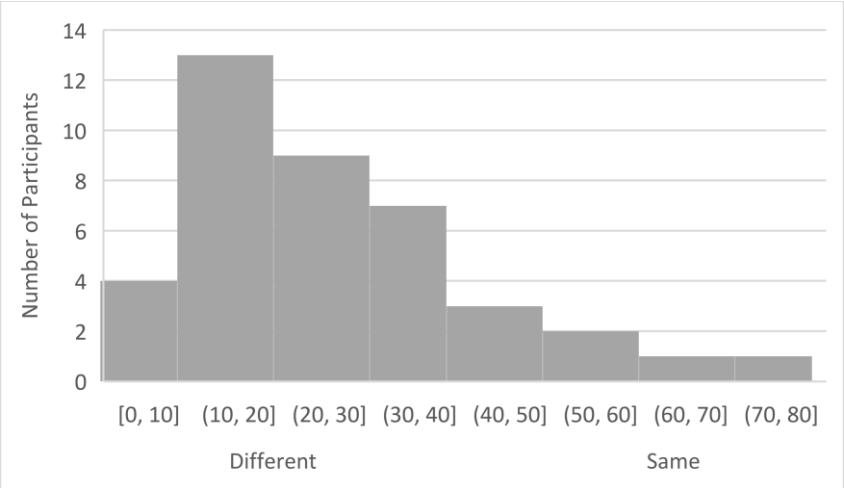

Figure 1. Distribution of Ratings of Similarity/Differences

Reliability

A trained investigator remeasured 20% of participants' visual analog scale protocols to confirm inter-judge reliability, for a total of 8 comparisons. Agreement was calculated using Cohen's kappa, which resulted in near perfect agreement (k = 0.964).

DISCUSSION

The purpose of this study was to investigate the difference in perception between the ModelTalker voice and the Message Banking voice. The results of this study suggest that although the sentences in both voices were equally intelligible, the Message Banking voice, the individual's actual voice, was preferable perceptually to the version created by ModelTalker. As hypothesized, the voices were found to be substantially different by the participants.

The hypothesis that Voice B (Model Talker) would be found to be less intelligible than Voice A (Message Banking) was not supported by the intelligibility data. In fact, both voices were equally intelligible; no errors were found in the orthographic transcriptions. This finding is similar to those of Cote-Giroux et al. who found that certain synthetic voices were statistically as intelligible as a natural human voice (2011). However, in that study the voices that were found to be as intelligible as human voices were also rated similarly to the human voice on perceptual measures, which was not the case in this study. A possible explanation for this finding is that the length of the sentences was short (5 words) and the sentences were grammatically simple (simple declarative and questions). In terms of complexity, the ability to correctly predict the content of the given sentences may have influenced accuracy of transcriptions. Longer sentences would also increase the demands put on the synthetic voice, as fluctuations in prosody increase in longer sentences. As noted by the participants, prosody was one of the more notable differences between the two voices and larger demands on the

voice may influence that further. Future studies should focus specifically on intelligibility with longer, more complex sentences to increase the validity and generalizability of the results, as it would better mimic real world usage of the voice. These results indicate that future studies are likely to find intelligibility differences, due to the differences in perceptual measures that were found in this study.

The hypothesis that there would be differences in perceptual parameters between the two voices was supported by the results of the data analysis. All three perceptual parameters were found to have significant differences between Voice A and Voice B. Naturalness had the largest mean difference and effect size, which was supported by results of the qualitative data. Naturalness was noted as an important factor by the majority of the participants, and the description of "robotic," especially, fits as naturalness being a significant difference between the two voices. Prosody was a quality that was noted by multiple participants, despite not being measured, or mentioned, in this study. Some participants specifically pointed out that the Message Banking voice had "natural prosody," indicating a possible relationship between prosody and naturalness. The relationship between prosody and the naturalness of synthetic voices has been an area of past research, especially in terms of the expression of emotions (Ilves & Surakka, 2012); however, prosody can also be considered a factor in pleasantness and clarity.

In addition to the limitations of utterance type noted above, there are several limitations of this study to consider. First, the demographics of the participants were not representative of people typically affected by ALS. This sample was homogenous in that all participants were female, young, Caucasian, and speech-language pathology students. These

disparities limit the generalizability of the results outside of this limited population. By nature of being speech pathology students, the participants very likely had prior knowledge and experience with both content and methodology, including the use of visual analog scales and the concept of rating vocal quality. The general population, including the people affected by ALS, are likely to have little or no experience with visual analog scales and voice quality rating, and thus, would be considered naïve participants.

Second, this population is very familiar with synthetic voices in everyday use, which can affect opinions of synthetic voices as noted by Delogu et al. (1998). Clients with less experience may have even less positive perspectives on the synthetic voices. It is likely that the general population, including people affected by ALS, will have more varied experience with technology. In the case of people with ALS, some participants may have more experience with synthetic voices due to personal experience with AAC or discussing options with doctors and therapists. As shown by Ball et al., people with ALS have a high rate of acceptance of AAC devices (2004). Richter et al. found that people with ALS preferred synthetic speech to natural, unintelligible speech (2009). Both of these results indicate people with ALS demonstrate an openness to AAC technology that may or may not be present in the sample population for this study.

An additional direction of research that may increase applicability and generalizability is the introduction of a professionally created synthetic voice as a potential factor for comparison. The use of a professionally created synthetic voice is a commonly used real world option for many people affected by ALS. Assessing preference between the personalized

synthetic voice, the professional synthetic voice, and the natural voice would better replicate the choice that most people with ALS face.

As noted above, an important direction for future research is to ask individuals with ALS and their caregivers to rate similar voices. Research conclusions based on data collected from peers may hold additional weight to affect decision making of people with ALS, which is the ultimate goal of this research.

Conclusion

Overall, this study was designed to provide information to allow clinicians to help their clients, especially those with ALS, make informed decisions about voicing choices for high tech AAC devices. It also indicated that there is a need for future research about this topic, including research that is more generalizable to the population that these results are intended to help.

References

Andersson, S., Yamagishi, J. & Clark, R. A. J. (2012). Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis. *Speech Communication, 54*, 175-188. doi: 10.1016/j.specom.2011.08.001

Ball, L. J., Beukelman, D. R., Anderson, E., Bilyeu, D. V., Robertson, J. & Pattee, G. L. (2007). Duration of AAC technology use by persons with ALS. *Journal of Medical Speech-Language Pathology, 15*(4), 371-381.

Ball, L. J., Beukelman, D. R. & Pattee, G. L. (2004). Acceptance of augumentative and alternative communication by persons with amyotrophic lateral sclerosis. *Augmentative and Alternative Communication, 20*(2), 113-122. doi: 10.1080/07434610420002165 96

Barsties, B. & De Bodt, M. (2014) Assessment of voice quality: Current state-of-the-art. *Auris Nasus Larynx, 42*, 183-188. doi: 10.1016/j.anl.2014.11.001.

Cote-Giroux, P., Trudeaus, N., Valiquette, C., Sutton, A., Chan, E. & Hebert, C. (2011). Assessment of nine French synthesized voices based on intelligibility and quality. *Canadian Journal of Speech-Language Pathology and Audiology, 35*(4), 300-311.

Delogu, C., Conte, S. & Sementina, C. (1998). Cognitive factors in the evaluation of synthetic speech. *Speech Communication, 24*, 153-168.

Drager, K. D. R., Reichle, J. & Pinkoski, C. (2010). Synthesized speech output and children: A scoping review. *American Journal of Speech-Language Pathology, 19*, 259-273.

Ilves, M. & Surakka, V. (2013). Subjective responses to synthesised speech with lexical

emotional content: The effect of the naturalness of the synthetic voice. *Behavior & Information Technology, 32*(2), 117-131. doi: 10.1080/0144929X.2012.702285

Kreiman, J. & Gerratt, B. R. (2000). Sources of listener disagreement in voice quality assessment. *Journal of Acoustic Society of America, 108*(4), 1867-1876.

Miller, R.G., Jackson, C.E., Kasarskis, E. J., England, J. D., Forshew, D., Johnston, W., … Woolley, S. C. (2009). Practice Parameter update: The care of the patient with amyotrophic lateral sclerosis: Multidisciplinary care, symptom management, and cognitive/behavioral impairment (an evidence-based review). *Neurology, 73*, 1227-1233

Mullenix, J. W., Stern, S. E., Wilson, S. J. & Dyson, C. (2003). Social perception of male and female computer synthesized speech. *Computers in Human Behavior, 19*, 407-424. doi: 10.1016/S0747-5632(02)00081-X.

Papadopoulos, K., Argyropoulos, V. S. & Kouroupetroglou, G. (2008). Discrimination and comprehension of synthetic speech by students with visual impairments: The case of similar acoustic patterns. *Journal of Visual Impairment & Blindness,* 420-429.

Richter, M., Ball, L., Beukelman, D., Lasker, J. & Ullman, C. (2003). Attitudes toward communication modes and message formulation techniques used for storytelling by people with amyotrophic lateral sclerosis. *Augmentative and Alternative Communication, 19*(3), 170-186. doi: 10.1080/0743461031000116544.

Stern, S. E., Mullennix, J. W. & Wilson, S. J. (2002). Effects of perceived disability on persuasiveness of computer-synthesized speech. *Journal of Applied Psychology, 87*(2), 411-417. doi: 10.1037//0021-9010.87.2.411

APPENDICES

Appendix A: Demographic Questionnaire

How old are you?

Have you ever been diagnosed with a severe hearing impairment? Yes No

How familiar are you with synthetic voices?
Unfamiliar Casual Familiar Knowledgeable Expert

In which of the following situations have you been exposed to a synthetic voice?
Cellphone Bank Computer GPS Video Game
Other: _____

Appendix B: Visual Analog Scale for Perceptual Characteristics

Phrase: _____

Unclear _____ Clear

Harsh _____ Pleasant

Synthetic _____ Natural

Appendix C: Visual Analog Scale for Overall Similarity

Overall, how would you compare the two voices?

Different _____ Similar

Appendix D: Grandfather Passage

You wish to know all about my grandfather. Well, he is nearly 93 years old, yet he still thinks as swiftly as ever. He dresses himself in an old black frock coat, usually several buttons missing. A long beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice is just a bit cracked and quivers a bit. Twice each day he plays skillfully and with zest upon a small organ. Except in the winter when the snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, "Banana oil!" Grandfather likes to be modern in his language.

Appendix E: Script for Conducting Study

Introduction:

"Hi everyone and thank you for agreeing to participate in my study. I am going to first hand out protocols and IRB consent forms. You do not have to sign a consent form because the IRB has determined that there is minimal risk in this study, but take a few moments to look over the form to familiarize yourself with the basics of the study. After you have finished, please fill out the demographic questionnaire on the first page of the protocol. Are there any questions?"

Before Starting Vocal Quality Portion:

"Has everyone completed their demographic questionnaire? Turn to the next page in your protocol, so that you can see the visual analog scales that you will be using in this study. You will be listening to short audio clips of sentences in two voices. Then you will write the sentence you heard on the line provided on your protocol. Next, you will rate the voices on pleasantness, clarity, and naturalness using a visual analog scale. All you need to do for the scale is to mark the line where you feel that the audio sample falls on the spectrum between

the two qualities written down. For example, if you feel the voice is very natural you would mark closer to the natural end of the line, and if you feel the voice is very unnatural you will mark your line towards the other end."

Before Starting Similarity/Difference Portion:

"Next you are going to hear longer passages in the same voices. You are going to rate these two voices on the visual analog scale on your last page. One end is same and the other is different.

Debriefing:

"One of the voices you heard was a recording of someone's natural voice. The other was a synthetic voice that was created to sound like the natural voice. This process is used by people who lose the use of their voice for many reasons, including ALS. Creating the synthetic voice you heard took eight hours of recording, and it could take even longer for someone with a disability. However, it can make an unlimited number of utterances, just like a professionally made synthetic voice. An alternate approach is to create recordings of your natural voice which is called message banking. This may take less dedicated time, but it results in a limited number of messages. There is a section for any comments you may have about the study like: Did you think the voices were similar or different? Why do you think the voices sounded different? Which voice did you prefer? Based on what I have told you about Message Banking and ModelTalker which method would you prefer?"

Appendix F: IRB Approval Letter

RESEARCH INTEGRITY AND COMPLIANCE
Institutional Review Boards, FWA No. 00001669
12901 Bruce B. Downs Blvd., MDC035 • Tampa, FL 33612-4799
(813) 974-5638 • FAX (813) 974-7091

April 10, 2017

Katherine Overton
Communication Sciences and Disorders
Tampa, FL 33612

RE: **Exempt Certification**
IRB#: Pro00029780
Title: Perceptual difference between natural speech and personalized synthetic speech

Dear Ms. Overton:

On 4/7/2017, the Institutional Review Board (IRB) determined that your research meets criteria for exemption from the federal regulations as outlined by 45CFR46.101(b):

(2) Research involving the use of educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures or observation of public behavior, unless: (i) information obtained is recorded in such a manner that human subjects can be identified, directly or through identifiers linked to the subjects; and (ii) any disclosure of the human subjects' responses outside the research could reasonably place the subjects at risk of criminal or civil liability or be damaging to the subjects' financial standing, employability, or reputation.

As the principal investigator for this study, it is your responsibility to ensure that this research is conducted as outlined in your application and consistent with the ethical principles outlined in the Belmont Report and with USF HRPP policies and procedures.

Please note, as per USF HRPP Policy, once the Exempt determination is made, the application is closed in ARC. Any proposed or anticipated changes to the study design that was previously declared exempt from IRB review must be submitted to the IRB as a new study prior to initiation of the change. However, administrative changes, including changes in research personnel, do not warrant an amendment or new application.

Given the determination of exemption, this application is being closed in ARC. This does not limit your ability to conduct your research project.

We appreciate your dedication to the ethical conduct of human subject research at the University of South Florida and your continued commitment to human research protections. If you have any questions regarding this matter, please call 813-974-5638.

Sincerely,

John Schinka, Ph.D., Chairperson
USF Institutional Review Board