Intelligibility and Attitudes Toward a Speech Synthesizer Vocoded Using Dysarthric Vocalizations

Rupal Patel, Ph.D.

Anna Roden, B.S.

Department of Speech Language Pathology and Audiology
Northeastern University, Boston, Massachusetts

Speakers with dysarthria who use augmentative and alternative communication (AAC) aids may rely on text-to-speech (TTS) synthesis as their means of interaction. Despite a variety of voices, clinicians tend to use the adult male voice given its superior intelligibility. While this practice may facilitate effective communication, it may come at the expense of naturalness and social acceptability, especially for children who use AAC. We have developed a prototype synthesizer that incorporates the residual prosodic cues in dysarthric vocalizations to convey voice identity. To compare intelligibility of a modified and unmodified synthesizer, we recruited 16 adults and 55 typically developing peers between 7 and 12 years of age to complete a listening task in which each synthesizer produced directions for navigating through a map of object landmarks. Additionally, adult participants assessed the age, gender, and naturalness of the voices, while child participants completed a survey to assess attitudes toward each voice. Results indicated that while the modified voice was significantly less intelligible, adult listeners correctly perceived the target user's age and gender. Children's intelligibility scores for both voices were considerably lower than those for adults, and responses on the attitude survey revealed that intelligibility was closely tied with preference. Trade-offs among intelligibility, acceptability, and preference are discussed.

Many children with severe dysarthria must rely on text-to-speech (TTS) synthesis as a primary means of interaction. Current commercially available augmentative and alternative communication (AAC) aids utilize a finite set of voices that are known to be relatively effective in everyday listening conditions such as classrooms and workspaces (Higginbotham et al., 1994; Logan, Green, & Pisoni, 1989). Although users can choose from a number of voices, many clinicians tend to program the device with an adult male voice given its superior intelligibility (Gorenflo, Gorenflo, & Santer, 1994; Greene, Logan, & Pisoni 1986; Higginbotham et al., 1994; Mirenda & Beukelman, 1987, 1990; Scherz & Beer, 1995). Studies on adult listeners report mean intelligibility scores that range between 81.2–96.7% for the most popular and most intelligible adult male voice—DECTalk Perfect Paul (Mirenda & Beukelman, 1987, 1990; Scherz & Beer, 1995). Female and child voices tend to be consistently less intelligible than male voices across a range of TTS types (Logan et al., 1989; Mirenda & Beukelman, 1987; Scherz & Beer, 1995). Children also find the adult male voice to be most intelligible; however, their scores tend to be more variable (35.3–92.9%) (Mirenda & Beukelman, 1987, 1990). While programming devices with the adult male voice may facilitate effective communication, it may come at

Journal of Medical Speech-Language Pathology
Volume 16, Number 4, pp. 243–249

the expense of naturalness and social acceptability, especially for children who use AAC. To date, this trade-off between intelligibility and uniqueness has received little attention in the literature.

The desire for natural and individualized speech output is evident in the literature. Nass and Lee (2000) found that in the general population, listeners tended to prefer TTS voices that resembled their own personality traits in terms of introversion and extroversion. Similarly Crabtree, Mirenda, and Beukelman (1990) studied age and gender preferences for synthetic speech in which they asked listeners to choose the voice they would want for themselves and found that the most natural-sounding gender-appropriate voices received the highest scores. Gorenflo et al. (1994) also found that listener attitudes were more favorable when nonspeaking individuals used TTS voices that closely resembled them in age and gender as compared to the individuals who attempted to communicate using unintelligible vocalizations.

Peer attitudes toward children who use AAC are influenced by a number of factors. O'Keefe, Brown, and Schuller (1998) found that peers were more likely to respond positively when an AAC user's message reflected their intelligence, age, and gender. Beck and colleagues have studied the effects of peer age, gender, and exposure to disabilities on attitudes. While Beck and Dennis (1996) found no differences in attitudes between children who attended inclusive schools versus those who did not, peer attitudes in inclusive schools became less favorable as children grew older (Beck, Fritz, Keller, & Dennis, 2000). Although peer attitudes did not differ for electronic versus nonelectronic aids, Beck, Bock, Thompson, and Kosuwan (2002) found that girls' mean scores tended to be more positive than the boys' (similar to Rosenbaum, Armstrong, & King, 1986). Lilienfeld and Alant (2002) also found more positive attitudes in girls; however, their results indicated attitudes were more positive toward users of devices with voice output compared to without. The challenge in developing TTS output for AAC aids appears to rest on balancing the demands of intelligibility, naturalness, and individualization.

In our laboratory, we are working toward developing an adaptive TTS synthesizer that incorporates the residual prosodic cues in dysarthric vocalizations to convey voice identity cues such as age, gender, and personality while maintaining intelligibility. The present study had three specific aims:

1. to compare intelligibility of our beta version of a modified voice with an unmodified voice,
2. to determine if the modified voice effectively conveyed user age and gender,
3. to determine whether peer attitudes would be more favorable toward a child who used the modified TTS voice rather than a child who used the unmodified voice.

METHODS

Participants

Sixteen adult speakers (8 males, 8 females, mean age = 22 years) of American English with normal hearing and no known speech, language, or cognitive impairment were recruited. Additionally, 55 typically developing 7- to 12-year-olds (22 males, 33 females, mean age = 9.4 years) with normal hearing function served as peer listeners.

Modified TTS Voice

Current TTS technology is limited in its ability to emulate the natural prosodic fluctuations of the human voice. Prosodic aspects of speech are not only essential for conveying linguistic differences; they also carry a majority of the cues to speaker identity. Vocal quality and prosodic cues such as pitch, loudness, and duration signal a range of speaker identity characteristics such as size, strength, age, gender, race, intellectual ability, attractiveness, and personality (c.f. Bachorowski & Owren, 1999; Collins, 2000; Hartman & Danhauer, 1976; Walton & Orlikoff, 1994; Zuckerman & Miyake, 1993). While speakers with dysarthria may exhibit inadequate articulatory control for producing speech sound segments, recent studies suggest that many individuals have preserved ability to vary prosodic features (Ciocca, Whitehill, & Ma, 2004; Le Dorze, Ouellet, & Ryalls, 1994; Patel, 2002, 2003, 2004). Our approach is to harness these residual abilities in order to emulate the vocal identity of the AAC user while maintaining intelligibility. As a first step toward this goal, we built a prototype voice for a 9-year-old boy with severe spastic dysarthria secondary to cerebral palsy, using vocoding techniques. An inventory of sustained-vowel vocalizations were gathered from the boy and used as the carrier signal. A set of 20 sentences were synthesized using a freeware concatenative synthesizer (Free TTS, Walker, Lamere, & Kwok, 2001) which

served as the modulator signal. The adult male voice (Kevin 16) was used as the default unmodified voice given its high intelligibility. The output of the synthesizer was then vocoded (Reason V3.0) using a subset of the dysarthric samples to produce the resultant modified utterances.

Stimuli

The synthesis process resulted in 20 unmodified sentences and 20 corresponding modified sentences, which served as stimuli for the intelligibility task. The stimuli consisted of directions for navigating through a map of object landmarks (referred to hereafter as the map task; Brown, 1995). Since the task was visually based and did not rely on the participant's reading/linguistic abilities, it could be completed by children and adults. The map specified start and finish locations and a number of object landmarks that were randomly dispersed between them. The background and the objects on the map were brightly colored to peak the participants' interest. To avoid ceiling effects in intelligibility, landmark names were minimal pairs, or rhyming words differentiated by only one sound (i.e., bat - hat).

Procedures

Listeners were provided with one of two maps. For each map, 10 instructions were provided using either the unmodified or modified voice to guide listeners from the start to finish destinations (see Figure 1 for an example). Each instruction ranged from 7 to 16 words in length. For example:

1. Draw a line from the dog to the cane.
2. From the cane, draw a line between the bun and the pin to the peas.

The length and complexity of the instructions was balanced across the two versions of the map. The order of the maps and the TTS voices were counterbalanced across participants.

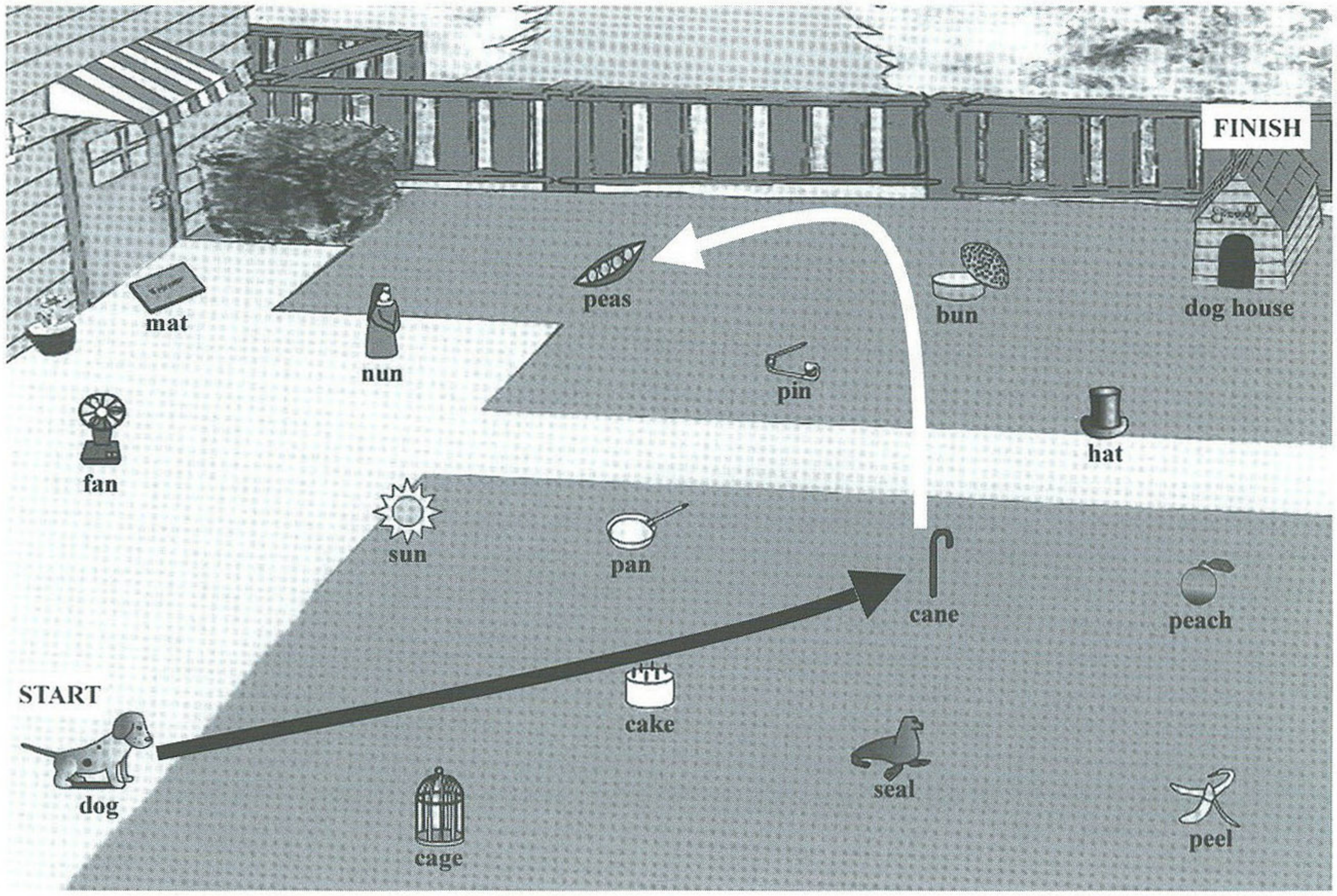

Figure 1. One of the maps used to assess intelligibility. The arrows indicate paths drawn by a listener in response to the first two instructions provided by the TTS.

While TTS intelligibility scores for both adults and children were based on map task performance, there were differences in the listening procedures across groups. Adults completed the listening experiment individually in a quiet room within our laboratory. Recordings were played through a computer using Windows Media Player and were presented over loudspeakers at an average sound pressure level of 70 dB. An interstimulus interval of 20 seconds was provided between instructions to allow participants sufficient time to process and respond to the directives. Upon completion of each map, adults were probed via questionnaire to estimate the age, identify the gender, and rate the quality of each voice on a five point scale (1 = very synthesized, 2 = synthesized, 3 = neutral, 4 = human, 5 = very human).

Data for the children were collected within the context of a classroom activity with group sizes varying from 4 to 16 children per classroom. Participants viewed two videos in which the target 9-year-old boy with cerebral palsy used an AAC device to provide map instructions. In video A, the AAC user communicated via the modified voice; in video B, he used the unmodified voice. Each video depicted a front view of the boy sitting in his wheelchair and interacting with his device to produce each instruction. This view was necessary for participants to gain a sense of the user's age, gender and personality as deemed though his body language and facial expressions. Although the same AAC user appeared in both videos (wearing different shirts) participants were told that the boys were twins in order to conceal the independent variable and to control for confounding variables such as physical appearance, attractiveness, and so forth, which have been noted to influence peer attitudes (c.f. Beck et al., 2000; Griffin & Langlois, 2006; Lilienfeld & Alant, 2002; Rosenbaum et al., 1986). A television and DVD player were used to present each video in the classroom, and the volume was adjusted to a comfortable hearing level based on participant feedback. An interstimulus interval of 40 seconds was provided between instructions to allow children to process and respond to each directive. Immediately following each video, children completed an adapted version of a 5-point (1 = strongly agree, 2 = agree, 3 = I don't know, 4 = disagree, 5 = strongly disagree), 37 item attitudes survey, the Communication Aid/Device Attitudinal Questionnaire (CADAQ) (Lilienfeld & Alant, 2002), which assessed the impact of the voices on peer perception along three dimensions (affective behavioral, cognitive competence, and communicative competence). This scale was chosen because it was the most applicable attitude survey for the age group and task at hand. Since the original CADAQ referenced two conversational partners, and the present study included only one communicator, some probes required slight rewording and one item was removed. Similar to Lilienfeld and Alant (2002), each item on the CADAQ was read aloud to the participants to control for reading abilities and to ensure participants stayed on task and answered all questions.

Data Analyses

A repeated measures group design was used to study the impact of TTS type on intelligibility in children and adults and on peer attitudes. Intelligibility scores were obtained from performance on the map task. Listener accuracy was defined in terms of word intelligibility, with a total of 26 points per map. Each of the 10 map directions was allocated between two and four points corresponding to target words. For example in "From the cane, draw a line between the bun and pin to the peas," four points could be earned for starting at "cane," ending at "peas," and drawing a path between "bun" and "pin." For each listener, a percentage accuracy score was calculated for each voice type. For the adults, a paired *t*-test was used to compare intelligibility of each TTS across participants at an alpha level of 0.05. For the children, a repeated measures ANOVA was performed with one within subjects factor (TTS type; modified vs. unmodified) and one between subjects factor (age group; 7–9 years vs. 10–12 years).

For the post-hoc questionnaire completed by the adults, age estimates were averaged across all participants. A paired *t*-test was used to compare the naturalness scores for each voice across participants at an alpha level of 0.05.

Peer attitudes toward each TTS voice were examined using the CADAQ. The 36 items on the scale were subdivided into 17 positive and 19 negative items. Following Oppenheim's (1973) procedures, scores were assigned such that high scores indicated more positive responses and low scores indicated more negative responses (for example, for positive items: 1 = strongly disagree, 2 = disagree, 3 = I don't know, 4 = agree, 5 = strongly agree, and for negative items: 1 = strongly agree, 2 = agree, 3 = I don't know, 4 = disagree, 5 = strongly disagree). A repeated measures ANOVA was per-

formed with one within subjects factor of TTS type (modified or unmodified) and two between subjects factors of age level (7–9 years or 10–12 years) and gender (male vs. female participants).

RESULTS

For the 16 adult participants, there was a statistically significant difference ($p < 0.001$) between average word intelligibility of the unmodified TTS (91.54%) compared to the modified TTS (68.85%) (Figure 2). Results of the ANOVA also revealed statistically significant differences between voices for the 55 child participants (F = 300.98; DF = 1, 53; $p < 0.0001$); average word intelligibility for the unmodified TTS was 65.19% versus 29.31% for the modified TTS (see Figure 2). There was no difference in intelligibility, however, between younger (7- to 9-year-olds) and older (10- to 12-year-olds) children ($p = 0.33$).

All 16 adults accurately identified the gender of the unmodified voice as an adult male, while 15 out of 16 participants judged the modified voice to be that of a young boy. The average age estimate for the unmodified TTS was 35.72 years compared to 9.19 years for the modified TTS. Notably, there was no difference in average naturalness scores between the unmodified TTS (1.54) and the modified TTS (1.64) ($p = 0.75$).

In terms of peer attitudes, ANOVA analyses revealed significant main effects of TTS type (F = 34.69; DF = 1, 51; $p < 0.0001$) and age group (F = 14.03; DF = 1, 51; $p < 0.001$) but not gender. None of the two-way and three-way interactions between TTS, age, and gender were statistically significant. CADAQ scores were more positive for the unmodified TTS (mean = 117.7) than the modified TTS (mean = 106.4) (note: higher scores reflect a more positive attitude). Additionally, older children (unmodified mean = 127.7; modified mean = 114.6) were more positive than younger children (unmodified mean = 109.4; modified mean = 99.5).

DISCUSSION

Results of the map task indicated that adults and children found the unmodified adult male voice to

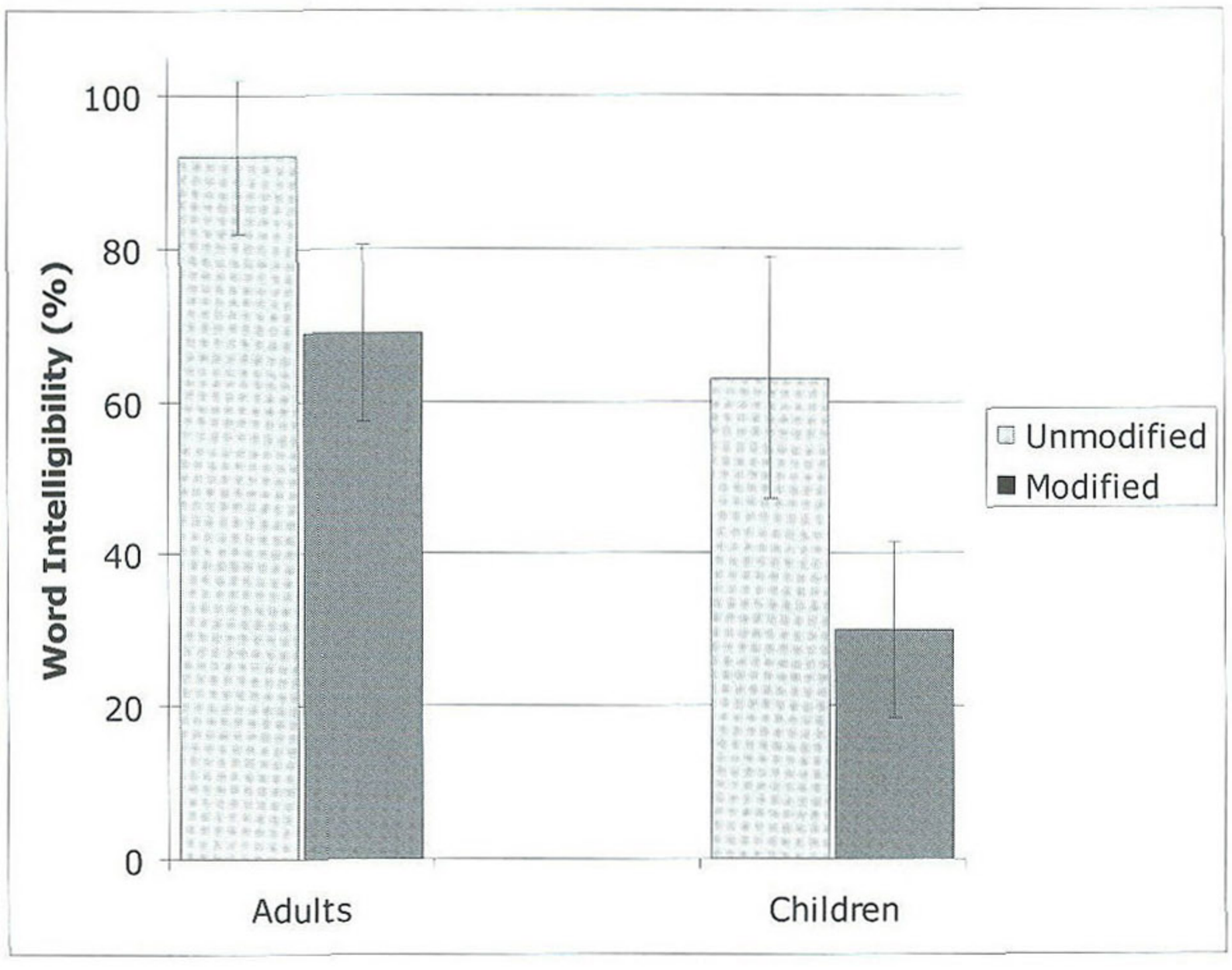

Figure 2. Average intelligibility scores of the modified and unmodified TTS across adults and children.

be more intelligible than the modified voice. While the trend in the present data is similar to that reported in the literature, our accuracy scores were lower for both TTS voices across adults and children (for example, see Logan et al., 1989; Mirenda & Beukelman, 1990; Scherz & Beer, 1995). Differences in accuracy across studies could be accounted for by differences in TTS voice type, data collection settings (lab vs. classroom; free-field listening vs. headphones), presentation mode (live vs. taped), and audio-visual equipment. Within the present study, adults heard stimuli over high quality computer speakers in a quiet room and were seated no more than 10 feet from the sound source, while the children heard stimuli in a large classroom via the audio routed out of a DVD player to the speakers on a television. The children were seated at their desks, which were commonly more than 10 feet from the sound source, and the level of ambient background noise fluctuated based on other activities within the school. While intelligibility of the modified voice in the laboratory setting may be acceptable, it degraded considerably in the classroom setting. Venkatagiri (2004) noted a similar loss in intelligibility when listening to synthesized speech in reverberant conditions. Future iterations of the modified TTS aimed at addressing the issue of room acoustics would be warranted.

The results were mixed in terms of perceptions toward the modified TTS. While adult participants accurately estimated the target age and gender of the modified voice, peer attitudes were less favorable toward this voice. In post-hoc discussions, many children commented that the modified TTS sounded like a "kid" and was a better voice match for the child in the video. When comparing attitudes toward the AAC user who used modified versus unmodified voice, the notion of a better match was largely overshadowed by the noticeable decrease in intelligibility. In fact, several children verbally expressed discontent while trying to follow the directives given by the modified TTS, and their frustration was reflected in poorer CADAQ scores. As Mirenda, Eicher, and Beukelman (1989) noted, perhaps adults put greater emphasis on naturalness while children place greater weight on intelligibility. Moreover, in the present study, children's attitudes may have been skewed toward intelligibility given the demands of the map task and its ordering with the attitude survey. If children believed poor performance on the map task reflected their own abilities, they may have responded more negatively to the less intelligible voice. Previous work has examined either intelligibility or attitudes but not the combination (Higginbotham et al., 1994; Lilienfeld & Alant, 2002). Given the multidimensional nature of peer attitudes, it is not surprising that TTS intelligibility influenced attitude responses.

In contrast to Beck et al. (2000), we found a significant difference in attitudinal responses between age groups, with the older group responding more positively than the younger group. Perhaps older children were more self-assured in their abilities to accurately follow directions and thus their performance on the map task did not influence attitude scores to the same extent as for younger children. Also contrary to previous work (Beck & Dennis, 1996; Beck et al., 2000; 2002; Lilienfeld & Alant, 2002; Rosenbaum et al., 1986), we did not find an attitude difference between boys and girls.

Despite poor intelligibility for the current version of the modified synthesizer, these initial findings are encouraging in that adults and children (noted anecdotally) were able to judge the target age and gender. Additionally, while the children's attitude scores were lower for the modified voice, they were comparable to those reported by Lilienfeld and Alant (2002). Novel methods for TTS adaptation that take into consideration source and filter characteristics of the target user to gain improved intelligibility and naturalness are currently underway. Just how much intelligibility would a young AAC user be willing to sacrifice for a voice that reflected her own unique identity? How much intelligibility loss would peers find tolerable? These questions are at the heart of the intelligibility-uniqueness trade-off and require further inquiry.

Acknowledgments This study was based on a Masters thesis completed by the second author, under the supervision of the first author. We are grateful to Michael Everett for implementing the modified synthesizer, to Matthew Fortier and Drew Rothstein for assistance in developing the audio and video stimuli, and to Pamela Campellone for her assistance with data collection. Finally, sincere thanks to the participants, the Boston Public School District, Brookline School District, and the Cranston-Johnson Catholic Regional School. This work was funded in part by the National Science Foundation Grant # 0712821.

Address correspondence to Rupal Patel, Ph.D., Department of Speech-Language Pathology and Audiology, Northeastern University, 360 Huntington Ave, 102 Forsyth Building, Boston, MA 02115 USA.
e-mail: r.patel@neu.edu

REFERENCES

Bachorowski, J. & Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *Journal of the Acoustical Society of America, 106*(2), 1054–1063.

Beck, A. R., Bock, S., Thompson, J. R., & Kosuwan, K. (2002). Influence of communication competence and AAC technique on children's attitudes toward a peer who uses AAC. *Augmentative and Alternative Communication, 18*, 217–227.

Beck, A. R., & Dennis, M. (1996). Attitudes of children toward a similar-aged child who uses augmentative communication. *Augmentative and Alternative Communication, 12*, 78–87.

Beck, A. R., Fritz, H., Keller, A., & Dennis, M. (2000). Attitudes of school-aged children toward their peers who use augmentative and alternative communication. *Augmentative and Alternative Communication, 16,* 13–26.

Brown, G. (1995). *Speakers, listeners and communication: Explorations in discourse analysis.* New York: Cambridge University Press.

Ciocca, V., Whitehill, T. L., & Ma, K-Y. J. (2004). The impact of cerebral palsy on the intelligibility of pitch-based linguistic contrasts. *Journal of Physiological Anthropology and Applied Human Science, 23*(6), 283–287.

Collins, S. A. (2000). Men's voices and women's choices. *Animal Behavior, 60*(6), 773–780.

Crabtree, M., Mirenda, P., & Beukelman, D. R. (1990). Age and gender preferences for synthetic and natural speech. *Augmentative and Alternative Communication, 6,* 256–261.

Gorenflo, C. W., Gorenflo, D. W., & Santer, S. A. (1994). Effects of synthetic voice output on attitudes toward the augmented communicator. *Journal of Speech Language and Hearing Research, 37,* 64–68.

Greene, B. G., Logan, J. S., & Pisoni, D. B. (1986). Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. *Behavior Research Methods, Instruments & Computers, 18*, 100–107.

Griffin, A. M., & Langlois, J. H. (2006). Stereotype directionality and attractiveness stereotyping: Is beauty good or is ugly bad? *Social Cognition, 24*(2), 187–206.

Hartman, D. E., & Danhauer, J. L. (1976). Perceptual features of speech for males in four perceived age decades. *Journal of the Acoustical Society of America, 59*(3), 713–715.

Higginbotham, D. J., Drazek, A. L., Kowarsky, K., Scally, C., & Segal, E. (1994). Discourse comprehension of synthetic speech delivered at normal and slow presentation rates. *Augmentative and Alternative Communication, 10*(3), 191–202.

Le Dorze, G., Ouellet, L., & Ryalls, J. (1994). Intonation and speech rate in dysarthric speech. *Journal of Communication Disorders, 27*(1), 1–18.

Lilienfeld, M., & Alant, E. (2002). Attitudes of children toward an unfamiliar peer using an AAC device with and without voice output. *Augmentative Alternative Communication, 18*, 91–101.

Logan, J. S., Greene, B. G., & Pisoni, D. B. (1989). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America, 86*, 566–581.

Mirenda, P., & Beukelman, D. R. (1987). A comparison of speech synthesis intelligibility with listeners from three age groups. *Augmentative and Alternative Communication, 3,* 120–128.

Mirenda, P., & Beukelman, D. R. (1990). A comparison of intelligibility among natural speech and seven speech synthesizers with listeners from three age groups. *Augmentative and Alternative Communication, 6,* 61–68.

Mirenda, P., Eicher, D., & Beukelman, D. R. (1989). Synthetic and natural speech preferences of male and female listeners in four age groups. *Journal of Speech Language and Hearing Research, 32*, 175–183.

Nass, C., & Lee, K. M. (2000). Does computer-generated speech manifest personality? An experimental test of similarity-attraction. *CHI Letters, 2*(1), 329–336.

O'Keefe, B. M., Brown, L., & Schuller, R. (1998). Identification and rankings of communication aid features by five groups. *Augmentative and Alternative Communication, 14*, 37–50.

Oppenheim, A. N. (1973). *Questionnaire design and attitude measurement.* London: Printa.

Patel, R. (2002). Phonatory control in adults with cerebral palsy and severe dysarthria. *Alternative & Augmentative Communication, 18,* 2–10.

Patel, R. (2003). Acoustic characteristics of the question-statement contrast in severe dysarthria due to cerebral palsy. *Journal of Speech, Language, & Hearing Research, 46*(6), 1401–1415.

Patel, R. (2004). The acoustics of contrastive prosody in adults with cerebral palsy. *Journal of Medical Speech-Language Pathology, 12*(4), 189–193.

Rosenbaum, P., Armstrong, R., & King, S. (1986). Improving attitudes toward the disabled: A randomized controlled trial of direct contact versus kids-on-the-block. *Developmental and Behavioral Pediatrics, 7*, 302–307.

Scherz, J. W. & Beer, M. M. (1995). Factors affecting the intelligibility of synthesized speech. *Augmentative and Alternative Communication, 11*(2), 74–78.

Venkatagiri, H. (2004). Segmental intelligibility of three text-to-speech synthesis methods in reverberant environments. *Augmentative and Alternative Communication, 20*(3), 150–163.

Walker, W., Lamere, P., & Kwok, P. (2001). "Introduction," http://freetts.sourceforge.net

Walton, J. H., & Orlikoff, R. F. (1994). Speaker race identification from acoustic cues in the vocal signal. *Journal of Speech and Hearing Research, 37*, 738–745.

Zuckerman, M., & Miyake, K. (1993). The attractive voice: What makes it so? *Journal of Nonverbal Behavior, 17*(2), 119–135.