ABSTRACT: Purpose: Research comparing the intelligibility of human and synthesized speech among both young children and adults has indicated that synthesized speech results in a degrading of intelligibility. The purpose of this study was to compare speech intelligibility of high-probability sentences produced using DECtalk® Perfect Paul and live speech among monolingual English-speaking and bilingual children.

Method: Twenty typically developing children between the ages of 4;5 (years;months) and 6;2 participated. Ten children (6 boys and 4 girls) were monolingual English speakers. The remaining ten children (4 boys and 6 girls) were bilingual. Their reproduction of modeled sentences produced in English via live speech and synthesized speech were analyzed to determine the intelligibility of each speech type.

Results: The results indicated that for both groups of children, performance was significantly better in the live speech condition. Results further revealed that the intelligibility decrement for synthesized speech was substantially greater for bilingual speakers.

Clinical Implications: For bilingual children, even higher quality synthesized speech may present a greater challenge than it does for monolingual English-speaking children. Possible strategies to enhance intelligibility are discussed.

KEY WORDS: synthesized speech, intelligibility, augmentative and alternative communication, bilingual speakers, children
population. Thirty-six percent (36%) of all schoolchildren served through the Individuals with Disabilities Education Act (IDEA) come from families of diverse linguistic backgrounds (U.S. Department of Education, 2000).

Unfortunately, many public schools continue to lack the capacity to develop and sustain practices that address the needs of linguistically diverse children. For example, a recent statewide survey (Kohnert, Kennedy, Glaze, Kan, & Carney, 2003) reported that only 27% of public school professionals in Minnesota had received any direct training in their professional program related to serving culturally and linguistically diverse populations, yet almost half (45%) provided services to clients from three or more different language backgrounds. Given the increasing demand for instructional support for large numbers of ELLs, it is likely that teachers will increasingly rely on English-language software that uses synthesized speech to supplement classroom instruction.

Typically, synthesized speech applications are considered in the context of implementing electronic communication aids to persons who cannot produce intelligible speech consistently. They allow individuals with severe expressive communication disabilities to express themselves using a wide array of potential communicative partners. Synthesized speech output also allows communication production via telephone and in classroom settings (without the need for close listener proximity to interpret symbol selections). Children may learn augmentative and alternative communication (AAC) system symbols more quickly with speech output (Schlosser, Belfiore, Nigam, Blischak, & Hetzroni, 1996), and speech output may facilitate comprehension and language learning for a beginning AAC user (Sevcik & Romski, 2002).

Children from linguistically diverse backgrounds are being increasingly exposed to synthesized speech applications as listeners of their peers who use AAC systems. Harrison-Harris (2002) reported that AAC system users include a large and growing constituency of children from linguistically diverse backgrounds. Although synthesized speech provides many potential benefits to persons who use electronic communication aids, it has the potential disadvantage of being less intelligible than natural speech production. Because of the increasingly larger number of children from linguistically diverse backgrounds both listening to synthesized speech applications and producing speech using these applications in both educational software and electronic communication aids, it becomes important to examine the intelligibility of frequently used speech synthesis software among young children.

There is general agreement in the empirical literature that, among low-cost speech synthesis software, DECTalk® is among the most intelligible (Mirenda & Beukelman, 1987, for a review). In single-word contexts, DECTalk is 60%–80% intelligible to typical adult native English speakers (McNaughton, Fallon, Neisworth, Tod, & Weiner, 1994; Greene & Pisoni, 1988; Mirenda & Beukelman, 1987). In more contextually rich sentences, DECTalk (i.e., Perfect Paul) is between 73% and 97% intelligible to adult native English speakers (Drager & Reichle, 2001b; Mirenda & Beukelman, 1987). A number of investigators have reported the Perfect Paul option to be the most intelligible voice in DECTalk across a range of chronological ages (Logan, Greene, & Pisoni, 1989; McNaughton, Fallon, Tod, Weiner, & Neisworth, 1994; Mirenda & Beukelman, 1987). In addition to the actual synthesized voice used, the variability reported in the intelligibility of synthesized speech can be influenced by a number of factors, including (a) speaking rate, (b) competing background noise, (c) linguistic context, (d) listening practice/training, and (e) age of listener. Each of these areas will be discussed briefly here.

Venkatagiri (1991) examined the influence that varying speech rate and pitch had on synthesized speech intelligibility. Among adults that he studied, slowing syllable production from 201 syllables per minute (spm) to 139 spm significantly improved speech intelligibility. Higginbotham, Drazek, Kowarsky, Scally, and Segal (1994) found that adult listeners were able to summarize text more accurately when they heard paragraphs spoken at very slow rates (10 s between each word = 5.5 words per minute [wpm]) than at normal rates (140 wpm). However, very slow rates such as these may interfere with the naturalness of the message. Sutton, King, Hux, and Beukelman (1995) reported that older adults (M = 69 years of age) preferred speaking rates between 130 and 210 wpm. Participants judged rates below 130 wpm as too slow and rates above 210 wpm as too fast. Among younger participants (M = 23.5 years of age), rates below 150 wpm were judged to be too slow and rates above 220 wpm were too fast.

Competing background noise is another factor that appears to influence synthesized speech intelligibility. Reynolds, Bond, and Fucci (1996) sought to determine synthetic speech intelligibility in the presence of background noise for both native and nonnative English adult speakers. Participants listened to 32 sentence pairs in quiet and in environments with a +10 signal-to-noise ratio. DECTalk was less intelligible in noise than in the quiet environment for both groups; however, nonnative English speakers experienced significantly more difficulty understanding DECTalk in noise than did native English speakers.

Linguistic context can also influence synthesized speech intelligibility. It is likely that the linguistic redundancy of longer utterances may compensate for a lack of acoustic redundancy in the synthesized signal. Highly predictable sentences or the provision of topic cues (most often with adults as participants) also appear to result in an enhancement of synthesized speech intelligibility (Hoover, Reichle, Van Tasell, & Cole, 1987; Marics & Williges, 1988; Oshrin & Siders, 1987; Slowiaczek & Nusbaum, 1985). Drager and Reichle (2001b) reported a facilitating effect of context on discourse when compared to sentence presentation for both younger adult and elderly adult listeners.

McNaughton et al. (1994) examined the effect that practice had on child (6- to 10-year-olds) and adult performance with speech synthesizers that included DECTalk. Participants received practice listening to lists of words during repeated sessions. Half of the lists involved listening to the same repeated words across each of five sessions. The remaining lists involved listening to novel words not previously heard during each of five sessions. Experimenters compared speech intelligibility between the
“repeated” and “novel” conditions. Children did not perform differently between the two conditions with DECTalk stimuli. However, with stimuli presented via Echo speech synthesis software (Street Electronics Corporation, 1982) (that had been shown to result in far lower intelligibility performance than DECTalk), children performed significantly better with practice in the “repeated” condition. Adults performed significantly better with practice in the “repeated” condition for stimuli presented by each of the two speech synthesizers.

For most of the available literature addressing synthesized speech intelligibility, young adults have served as participants. Results from investigations that have specifically addressed the effect of age on intelligibility have been equivocal. Humes, Nelson, Pisoni, and Lively (1991) and Humes, Nelson, Pisoni, and Lively (1993) reported that age was not a significant factor on synthesized speech intelligibility. In comparing young and elderly adult performance, Drager and Reichle (2001b) reported no differences in speech intelligibility in sentences or discourse. On the other hand, Kangas and Allen (1990) concluded that age was a critical variable. Very few researchers have investigated the effect of synthesized speech on children’s listening performance. Mirenda and Beukelman (1987) examined the intelligibility of various synthesizers with both children (ages 6–8 and 10–12) and adults (ages 26–40) listeners. They demonstrated that children had greater difficulty than adults with the synthetic speech. Reynolds and Jefferson (1999) examined the ability of two groups of children (ages 6–7 and ages 9–11) to comprehend natural speech and DECTalk using a sentence verification task. They found that children’s intelligibility was lower than that found in previous investigations with adults. They also found that both groups of children performed significantly better with live speech than with DECTalk.

To date, relatively little attention has been given to the effect that chronological age and language experience may have on the performance of very young children. Instead, the bulk of the available research examining synthesized speech intelligibility has focused on adult native speakers of English. However, as established earlier, an increasing number of non-English speakers may benefit from an electronic communication aid with synthesized speech output. Previous research suggests that nonnative speakers may experience more difficulty deciphering synthesized speech than native English speakers (Mack, Tierney, & Boyle, 1990; Reynolds, Bond, & Fucci, 1996). Mack et al. (1990) reported that nonnative adult speakers made significantly more errors than native English speakers in identifying consonant-vowel-consonant (CVC) synthesized words. However, there were no significant differences between the two groups with stimuli presented in natural speech. The participants in this investigation came from extremely diverse linguistic backgrounds (Japanese, Chinese, Arabic, Korean, Russian, Indonesian, Spanish, Ewe, and Thai languages). Additionally, the degree of English experience of participants varied from 6 months to 23 years.

Rarely has the performance by young children and individuals from culturally and linguistically diverse backgrounds been addressed adequately in examining synthesized speech applications. Consequently, the purpose of the current investigation was to compare the intelligibility of sentences produced using DECTalk Perfect Paul and human-recorded speech among young children who are monolingual speakers of English as compared to their peers learning English as a second language.

**METHOD**

**Participants**

Participants included 20 children between the ages of 4;5 (years;months) and 6;2. Ten of these children (6 boys and 4 girls) were monolingual English-speaking children (mean age = 5.02 years). The remaining 10 children (4 boys and 6 girls) were bilingual (mean age = 5.20 years). The native languages for these participants included Sudanese (3 participants), Indonesian (2 participants), and Hmong (5 participants). All bilingual children were sequentially bilingual and were born in the United States. Their parents reported equal proficiency in both English and their non-English language. The bilingual participants’ primary language spoken at home was their native language. Children experienced preschool educational services beginning by 2 years of age. English was spoken exclusively in the children’s public school program. Table 1 includes participants’ characteristics.

Each child successfully passed a hearing screening at 20db HL at frequencies including .5k, 1K, 2K, and 4K. The experimenter implemented a pretest to ensure that each participant was able to perform the task required to obtain the dependent measure. The experimenter asked each participant to repeat each of the five sentences.

<table>
<thead>
<tr>
<th>Participant</th>
<th>Native language</th>
<th>Chronological age</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>English</td>
<td>4;10</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>English</td>
<td>4;6</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>English</td>
<td>4;8</td>
<td>F</td>
</tr>
<tr>
<td>4</td>
<td>English</td>
<td>5;10</td>
<td>F</td>
</tr>
<tr>
<td>5</td>
<td>English</td>
<td>5;0</td>
<td>M</td>
</tr>
<tr>
<td>6</td>
<td>English</td>
<td>4;8</td>
<td>M</td>
</tr>
<tr>
<td>7</td>
<td>English</td>
<td>6;0</td>
<td>M</td>
</tr>
<tr>
<td>8</td>
<td>English</td>
<td>5;0</td>
<td>M</td>
</tr>
<tr>
<td>9</td>
<td>English</td>
<td>4;5</td>
<td>M</td>
</tr>
<tr>
<td>10</td>
<td>English</td>
<td>5;4</td>
<td>M</td>
</tr>
<tr>
<td>11</td>
<td>Sudanese/English</td>
<td>5;0</td>
<td>F</td>
</tr>
<tr>
<td>12</td>
<td>Sudanese/English</td>
<td>6;0</td>
<td>F</td>
</tr>
<tr>
<td>13</td>
<td>Sudanese/English</td>
<td>5;0</td>
<td>M</td>
</tr>
<tr>
<td>14</td>
<td>Indonesian/English</td>
<td>5;1</td>
<td>F</td>
</tr>
<tr>
<td>15</td>
<td>Indonesian/English</td>
<td>5;6</td>
<td>M</td>
</tr>
<tr>
<td>16</td>
<td>Hmong/English</td>
<td>5;0</td>
<td>M</td>
</tr>
<tr>
<td>17</td>
<td>Hmong/English</td>
<td>6;1</td>
<td>F</td>
</tr>
<tr>
<td>18</td>
<td>Hmong/English</td>
<td>6;2</td>
<td>M</td>
</tr>
<tr>
<td>19</td>
<td>Hmong/English</td>
<td>5;5</td>
<td>F</td>
</tr>
<tr>
<td>20</td>
<td>Hmong/English</td>
<td>4;2</td>
<td>F</td>
</tr>
</tbody>
</table>
presented live and in English that matched the length and complexity of stimuli used in the experimental tasks. To qualify for participation in the experimental conditions, participants had to repeat each word successfully in a sentence with no more than one word deleted/ replaced for each of the five sentences. All children successfully completed this pretest with the exception of 1 Hmong child who did not participate in the experimental conditions.

Setting

Experimental sessions occurred in a quiet room of the children’s homes or in the office of the children’s church. Only the experimenter, participant, and reliability agent were present during experimental sessions. In a prior investigation conducted in very similar environments (Alamsaputra, 2004), the noise level had been established using a sound level meter (Qwest Model CA-22) to establish a measure of environmental noise. With the sound level meter set at 250 Hz reference and using a calibration tone at 94 dBA, the average noise level was 37 dBA (range 30–42 dBA).

Stimuli

Stimuli included 32 sentences that were randomly selected from the High Probability Sentences for Phonetic Inventory (SPIN) list developed by Kalikow, Stevens, and Elliott (1977). SPIN sentences were controlled for word predictability. The high-probability sentences had a predictable last word because of syntactic, semantic, and prosodic cues available in the sentences (e.g., Cut the bacon into strips). The sentences in the SPIN list were divided into eight lists consisting of 25 high-probability and 25 low-probability (e.g., Bob heard Paul called about the strips) sentences each. Thirty-two (32) high-probability sentences were chosen randomly from those eight lists and were divided into two new lists (List A and List B) consisting of 16 sentences each. Although typically used with older children, stimuli from the SPIN have been used in a large number of studies examining speech intelligibility (Clopper et al., 2000).

The experimenter presented the speech output for the sentences in the following manner: A female speaker with a midwestern dialect presented List A live. List B sentences were spoken by Perfect Paul from DECtalk using the DynaVox program (DynaVox, Inc.). All stimuli were presented via stereo speakers (Altec Lansing Multimedia Computer Speaker System ACS21W). The speakers were connected to a laptop computer (DELL Model PP01L).

Procedure

The experimenter presented all stimuli at a rate of 185 wpm. Participants established a comfortable listening level for both types of speech stimuli before the presentation of the sentences. The experimenter provided each participant with the following instructions:

You are going to listen to some sentences. You need to repeat exactly what I say the best that you can. Ready? Okay.

After the testing, the experimenter thanked each participant and offered a small toy. Each participant listened to the sentences in both natural and synthesized speech. Half of the participants from each group were presented with the human speech condition first, and the other half were presented with Perfect Paul first. Experimenters counterbalanced sentence presentation order across participants. After participating in the first condition, each participant received an opportunity for a break before listening to the remaining condition.

Data Analysis

The independent variables included group (monolingual English speakers and bilingual speakers) and speech type (human voice and synthesized speech [Perfect Paul]). The dependent measure consisted of correct lexical repetitions of words. To be considered exact, a word had to be identified by the experimenter as a word corresponding to the word on the printed sentence list of the SPIN sentence. Phonological variations that could be identified as an intelligible word were counted as exact lexical productions.

Each child-produced sentence repetition was audiorecorded and subsequently transcribed by the experimenter and a reliability agent. Investigators compared each word transcribed to the actual word that the child spoke. Each word spoken by the child had to lexically match that of the sample stimulus and appear in the same order as the word in the sample stimulus.

A multiple analysis of variance (ANOVA) for Group (monolingual and bilingual) × Condition (live and DECtalk) compared performance across the two groups sampled. Effect sizes were determined following procedures delineated by Cohen (1988).

Interrater Reliability

The primary author and a second independent observer computed interrater reliability for all utterances produced by 20% of the participants in both experimental conditions. The primary author and the reliability observer were graduate students in communication disorders and had extensive experience transcribing and coding child language samples. The experimenter did not conduct specific training before data scoring other than reviewing the definition of an exact response. Both raters transcribed an audio recording of child responses and scored the accuracy of the responses independently. Then, interobserver reliability using item-by-item agreement was computed. To be an agreement, the transcription and scoring of the child’s response between observers had to match. The experimenter computed reliability by dividing the agreements by the agreements plus disagreements and multiplying the outcome by 100. Reliability was 100% for transcription and judgment of accuracy for both groups of participants in each of the within-participant experimental conditions.
RESULTS

The percentage of total words correctly produced by monolingual English-speaking and bilingual children in human-recorded speech and DECTalk experimental conditions is shown in Figure 1. The total number of words possible was 105 words in the live voice condition and 106 words in the DECTalk condition. The percentage of words accurately repeated for the monolingual group was 99% (range = 97%–100%; SD = 1.0) in the live voice condition and 84% (range 73%–100%; SD = 6.0) in the DECTalk condition. For the bilingual group, the percentage of words accurately repeated was 92% (range = 83%–98%; SD = 6.3) in the live voice condition and 61% (range = 39%–93%; SD = 13%) in the DECTalk condition. Note that the standard deviations for children in the bilingual group were more than double that for the monolingual group in both conditions.

Percent word accuracy scores were entered into a 2 (Group) × 2 (Condition) mixed ANOVA. There were main effects of group, $F(1, 15) = 24.679, p < .001, \eta^2 = 2.3$, reflecting the greater number of correct words for monolingual as compared to bilingual, and condition, $F(1, 15) = 85.082, p < .001, \eta^2 = 3.0$, reflecting the greater performance for both groups in the live voice condition. Importantly, these main effects were qualified by a significant Group × Condition interaction, $F(1, 15) = 12.862, p < .001, \eta^2 = 3.0$. Figure 1 displays this interaction. Experimenters implemented a Bonferroni correction to lessen the possibility of a Type 1 error. Pairwise comparisons using Bonferroni correction indicated that the source of this interaction was the disproportionate decline in word accuracy across conditions for the bilingual group relative to the monolingual group. The monolingual group repeated 84% of the DECTalk words contained in sentence stimuli. That is, for bilingual children, average accuracy dropped by a significant 30.9% from live speech to DECTalk. Although accuracy for the monolingual group was also negatively affected in the DECTalk condition, this decline in performance was much less (15%). As indicated by the effect sizes, the magnitude of all group and condition differences was large.

DISCUSSION

Monolingual and bilingual children performed similarly in repeating sentences presented using live speech. For stimuli presented via DECTalk Perfect Paul, there was a significant difference in intelligibility between the two groups. Monolingual children found 84% of the DECTalk words contained in sentence stimuli intelligible. This percentage is comparable to that reported in previous investigations addressing DECTalk intelligibility among children (McNaughton et al., 1994; Mirenda & Beukelman, 1987). In the current study, bilingual children correctly repeated a mean of 61% of total words contained in sentences presented via DECTalk Perfect Paul. Perfect Paul’s intelligibility is apt to make it difficult for bilingual children to understand messages produced by a child who used an electronic communication aid using that speech synthesis software.

The current investigation used an environment that did not have a tight control for average noise level. Elementary classroom settings often have a significant level of background noise that can approximate 55–65 dBA (Nelson, Soli, & Seitz, 2002). Available literature suggests that a signal-to-noise differential of as small as 10 dB can result in significantly different intelligibility when listening to synthesized speech output. Future investigations should carefully examine the role that signal-to-noise ratios may have on the speech synthesis intelligibility for the two populations sampled in this investigation. Given previous work in similar environments (Alamsaputra, 2004), it is possible that the experimental environment may have been quieter than a typical classroom. If true, it is possible that the results of the current investigation actually yielded slightly better performance than would occur in a classroom environment. On the other hand, because contextual information was limited in the current investigation, it would be important to replicate the current study in a classroom environment where noise levels were held within a narrow range.

Children in the current investigation had no known previous exposure to DECTalk. There is evidence suggesting that synthesized speech intelligibility improves with practice. As discussed in the introduction, McNaughton et al. (1994) found that 6- to 10-year-olds’ performance with DECTalk improved significantly (69% to 86%) with repeated listening to single-word stimuli across five sessions. Additionally, this study demonstrated that performance across five sessions improved for children listening to novel words with which they had no previous synthesized speech experience (71% to 84%). The McNaughton et al. investigation addressed children who were native English speakers. It is unclear whether bilingual speakers
would demonstrate the same level of improved performance with practice involving repeated or novel stimuli. However, answering this question would help clarify the practical importance of the results obtained in the current investigation. If bilingual children show a steeper practice effect, the initial lower intelligibility would create an initial formidable communicative challenge. However, this challenge could be overcome quickly in a classroom setting.

**Future Research**

Future investigations are needed that examine the intelligibility and practice effect for other DECTalk voices. Previous investigations have demonstrated that Perfect Paul is the most intelligible DECTalk voice. With children, Kit the Kid may be the most likely choice of DECTalk voices because of its chronological appropriateness. However, if Kit the Kid is less intelligible than adult voice options available in DECTalk, interventionists could face a dilemma in choosing a voice. Future investigations should examine the intelligibility of Kit the Kid with EO and bilingual children in natural environmental contexts to examine the degree to which Kit the Kid degrades intelligibility performance when compared with an adult voice such as Perfect Paul.

A final area of investigation should involve the differentiation of intelligibility from comprehension of synthesized speech applications for young children. There are data suggesting that when elderly adults listen to DECTalk Perfect Paul in narrative presentations, comprehension (answering questions that require understanding) deteriorates significantly more than during a human-recorded speech condition when compared to reproduction of words (Drager & Reichle, 2001a). One explanation for this difference is that listeners may allocate so much effort to deciphering the message (intelligibility) that they divert cognitive resources from processing the actual message (Duffy & Pisoni, 1992).

**Clinical Implications**

Bilingual participants in the current investigation found synthesized speech to be challenging. However, a number of implications may be drawn from the current investigation. Interventionists should consider using contextual cues whenever possible to supplement the use of synthetic speech applications. In the current investigation, only linguistic redundancy was available as a contextual cue. Second, interventionists should consider using a speech rate for synthesized speech that is within the slower end of a normal range of speech rate.

Another important consideration is the noise level in the listening environment. When using synthesized speech in classroom settings, interventionists may need to consider the use of external speakers that are located near prospective listeners. Additionally, it may be necessary to consider strategies to reduce classroom noise levels in environments where there are children listening to synthesized speech applications. Finally, it is important that interventionists appropriately adjust their expectations for the possibility that synthesized speech applications will result in a degrading of learner performance when compared to their performance as a result of listening to live speech. There may be a tendency for interventionists to assume that learners are being less compliant when listening to synthesized speech when they may simply have difficulty receiving the synthesized output. Because teachers and teaching assistants represent primary organizers of educational environments, it is important that speech-language pathologists collaborate closely in considering strategies to maximize the intelligibility of classroom synthesized speech applications.

**SUMMARY**

Results from the current investigation suggest that bilingual children have significantly more difficulty deciphering synthesized speech than human-recorded speech. This difference was substantially greater than that found for a nonmatched monolingual English-speaking group of typically developing children with comparable chronological ages. With electronic communication devices becoming increasingly available in public school settings, it is important to examine the efficiency of synthesized speech output. Examining the relative intelligibility and comprehensibility of synthesized speech output represents an important contributor to the ease with which social interactions can be facilitated among AAC users and their peers.

**REFERENCES**


