JSLHR, Volume 41, 744–752, August 1998

DECTalk and MacinTalk Speech Synthesizers: Intelligibility Differences for Three Listener Groups

Katherine C. Hustad
Department of Special Education
and Communication Disorders
University of Nebraska, Lincoln

Ray D. Kent
Department of
Communicative Disorders
University of Wisconsin, Madison

David R. Beukelman
Department of Special Education
and Communication Disorders
University of Nebraska, Lincoln

This study examined word level intelligibility differences between DECTalk and MacinTalk speech synthesizers using the Modified Rhyme Test in an open format transcription task. Three groups of listeners participated: inexperienced, speech-language pathologists, and speech synthesis experts. Results for between-subjects ANOVA showed that the expert group correctly identified a significantly higher number of words than each of the other listener groups. For the within-subjects factor of voice, simple effects ANOVA and post hoc contrasts within each group showed that listeners had higher intelligibility scores for the DECTalk male voice, Perfect Paul, than for the MacinTalk male voice, Bruce. No other pairwise gender/age-matched differences were found between the two synthesizers.

KEY WORDS: speech synthesis, DECTalk, MacinTalk, intelligibility, listening experience

During the last two decades important advances have been made in augmentative/alternative communication (AAC) technology resulting in electronic communication systems that are more sophisticated and more easily accessed by people with disabilities. Improvements in the quality of voice output allow AAC users to communicate independently in a variety of communicative contexts. Many AAC systems housed in personal computers are now available with built-in software for synthesized speech. This has several advantages for the user including lower cost and increased portability.

At the present time, a formant-based speech synthesis system, DECTalk™ (Klatt, 1980; Klatt & Klatt, 1990), is the most widely used speech synthesizer in AAC technology. Studies have shown that DECTalk is the most intelligible speech synthesizer at the word level (Greene, Logan, & Pisoni, 1986; Logan, Greene, & Pisoni, 1989; Mirenda & Beukelman, 1987) and at the sentence level (Mirenda & Beukelman, 1987; Scherz & Beer, 1995). AAC developers have imported DECTalk voices into newer systems because of this high level of intelligibility.

The algorithm employed in DECTalk is based on detailed theoretical foundations from the acoustic theory of speech production (Fant, 1960). For example, DECTalk employs two different sound sources, one for voicing and one for noise. In addition, two sets of resonators are used, a serial configuration for vowels and a parallel configuration for

fricatives. In all, 39 different parameters are configured in DECTalk, and they are updated every 5 milliseconds. Extensive language-specific pronunciation rules as well as a dictionary of exceptions increase the likelihood that messages entered as text are spoken correctly. Ten different voices are produced by the DECTalk synthesizer. Each is a variant of the primary voice, Perfect Paul. Differences between the voices reflect variations in parameter specifications with each voice representing perceptual distinctions relating to gender, age, body size, and voice characteristics.

Recently, a new generation of software-based text-to-speech synthesis systems has been developed for personal computer-based AAC systems. For example MacinTalk™, a program developed by Apple Computer, is widely employed by AAC users who use a Macintosh platform.

MacinTalk employs diphone-based linear predictive coding (Venkatagiri, 1996). Diphone synthesis involves encoding different linguistic units than conventional formant synthesis. Whereas formant synthesis typically encodes individual phonemes, diphone synthesis incorporates phoneme boundaries. Diphones consist of segments extending from the steady state center of one phoneme to the steady state center of the next phoneme. As a result, coarticulatory information is encoded into the algorithm (Venkatagiri, 1996). Diphones are extracted from recordings of words within carrier phrases produced by a natural talker. Because an actual human voice is the source of the sound units, diphone synthesis is thought to be more natural sounding than formant synthesis, which is entirely machine generated (Mirenda & Beukelman, 1990). Diphone-based synthesis requires that all possible phoneme pair permutations occurring in the language be encoded to ensure pronunciation accuracy. Production of individual words and sentences involves concatenating the appropriate diphones. Like DECTalk, the MacinTalk synthesizer employs a dictionary of exceptions to improve pronunciation accuracy.

Diphone synthesizers require greater storage capacity and microprocessor power (Mirenda & Beukelman, 1990) than formant synthesizers. In formant synthesis, only individual phonemes need be stored whereas in diphone synthesis, all encoded sound pair permutations found in the language must be stored (Venkatagiri, 1996). In the past, computer capability issues were of concern. In general, open-format word-level intelligibility testing of earlier diphone-based synthesizers such as Real Voice and Smoothtalker showed that it was inferior to high quality formant synthesis (i.e., DECTalk) (Logan et al., 1989). However, as technology continues to advance, and memory and processing power become minor considerations, diphone synthesis holds renewed potential. In a study examining French diphone synthesis, O'Shaughnessy, Barbeau, Bernardi, and Archambault (1988) concluded that the intelligibility of diphone synthesizers could potentially surpass even high-quality formant synthesizers.

Several versions of MacinTalk have been released, each containing different voices. One of the earlier versions, MacinTalk Pro™, contained generic high quality male, female, and child voices. Later versions included the MacinTalk II Pro™ and the MacinTalk III Pro™, each of which have several different male, female, and child voices. The primary difference between later versions is that MacinTalk II Pro requires 8 megabytes of RAM, whereas MacinTalk III Pro, the most recent version, requires only 4 megabytes of RAM. Approximately 20 different voices varying in gender and vocal quality, similar to the different DECTalk voices, are available with the MacinTalk II and III Pro software modules.

To date, only one study has examined the intelligibility of MacinTalk voices. Rupprecht, Beukelman, and Vrtiska (1995) studied the differences between early versions of MacinTalk Pro male, MacinTalk Pro female, DECTalk Paul, DECTalk Betty, and the original MacinTalk synthesizer with inexperienced listeners. The experimental task involved transcribing the final word in each of a series of sentences taken from the Speech in Noise test (SPIN) (Kalikow, Stevens, & Elliot, 1977). Both high and low predictability sentences were employed, and data reported reflect means across both sentence types. There were no significant differences between individual MacinTalk Pro voices and DECTalk voices. However, the original MacinTalk voice was significantly worse than the others. Based on these results, Rupprecht and colleagues concluded that MacinTalk Pro and DECTalk voices were equivalent. Comparison of these results with intelligibility of other speech synthesizers is difficult because most existing data have not been obtained from SPIN stimuli. In addition, because linguistic context is present in half of the SPIN sentences, predictability of target words for those sentences is optimized. Studies of later versions, MacinTalk II and III Pro, have not been conducted.

Issues related to learning have also been of interest to researchers in the area of speech synthesis. In general, studies have shown that intelligibility scores improve with listener experience. Several studies have explored the extent to which it improves in as few as one (Venkatagiri, 1994) to as many as eight learning sessions (Schwab, Nusbaum, & Pisoni, 1983). Results have varied considerably depending upon the synthesizer, stimulus material, and teaching methods. McNaughton, Fallon, Tod, Weiner, and Neisworth (1994) found significant improvement between the first and the last of five learning sessions with the DECTalk child

voice, Kit, on an open response word level intelligibility task. Schwab et al. (1983) found that listeners who received training with the Votrax speech synthesizer showed improvement between each of eight learning sessions on word and sentence length material but not on paragraph length material. Further, they found that listeners who received training using synthesized speech performed better than those who received training using natural speech. In fact, the performance of listeners who received training using natural speech did not differ from the performance of listeners who received no training. Clearly, learning can have an important effect on intelligibility, and this effect relates to the functional usefulness of synthesized speech in communicative contexts.

There were two primary purposes for this study. First, this study sought to determine whether there are differences between DECTalk and later versions of the MacinTalk Pro synthesizers as a whole (i.e., across voices) and for gender/age-matched individual voices (male, female, child) within three groups of listeners on isolated word-level stimuli. Second, this study sought to determine whether or not there are differences in intelligibility scores among listeners with different kinds of listening experience across the two types of synthesizers. Listener experience groups included: inexperienced listeners, speech-language pathologists (SLP), and expert listeners of synthesized speech. Given previous research demonstrating the high intelligibility of the DECTalk synthesizer on isolated word-level stimuli and lack of data for the MacinTalk synthesizer on similar stimuli, it was hypothesized that DECTalk would be better than MacinTalk overall within each group. In addition, it was expected that individual DECTalk voices would be better than individual age/gender-matched MacinTalk voices within each level of experience. Based on the findings of Schwab et al. (1983), it was hypothesized that the group of experts would obtain significantly higher intelligibility scores across synthesizers than the SLPs and the inexperienced listeners. Because neither inexperienced listeners nor SLPs had more than incidental experience listening to speech synthesis, it was expected that both groups would have similar intelligibility scores.

Methods

Participants

A total of 18 individuals participated in this study, 6 in each of three groups. The first group consisted of inexperienced listeners, the second of SLPs, and the third of expert listeners. All participants reported no known neurological deficits, language/learning disabilities, or hearing loss. Audiological evaluations were not performed. No gender criteria were imposed on any group. Listeners were paid $10 for their participation, which required a one-time commitment of approximately 1 hour. Table 1 details listener characteristics by group. Participants within each group were selected on three sets of criteria.

1. *Inexperienced listener group.* Listeners in this group were required to have no more than incidental experience listening to synthesized speech. In addition, they did not have specialized knowledge of speech acoustics or speech perception. All listeners were native speakers of American English. Four participants held professional positions and 3 were currently enrolled in graduate school. The gender composition was 4 females and 2 males.

2. *Speech-language pathologists (SLP).* These participants were selected based on highly specialized skills as listeners. All participants in this group met the following criteria: (a) had at least 2 years of clinical experience as an SLP; (b) had studied speech acoustics in a doctoral-level course; (c) had no more than incidental experience listening to synthesized speech in the past 4

Table 1. Mean, standard deviation, and range for age and experience listening to synthesized speech for all listener groups.

Group	Mean (*SD*)	Range
Inexperienced listeners		
Age	31.0 years (9.86)	23–50 years
Speech synthesis experience	0	0
Speech-language pathologists (SLP)		
Age	33.3 years (5.46)	26–40 years
Speech synthesis experience	0	0
Expert listeners		
Age	39.5 years (9.00)	26–51 years
Years of experience with synthesized speech	6.0 years (3.56)	1.5–10 years
Hours per week for 6 months prior to this study	9.6 hours (3.67)	6–16 hours

years, and (d) were currently pursuing PhD studies in speech-language pathology on a full-time basis.

All listeners spoke American English as their primary language, and all but one were native speakers of American English. This non-native individual, however, had been speaking English as a primary language for the past 19 years, had been a resident of the U.S. for the past 15 years, and had received all post-secondary education in English. The gender composition of this group was 4 females and 2 males.

3. *Expert synthesized speech listeners.* Listeners in this group were selected based on extensive experience with synthesized speech. All participants in this group met the following criteria: (a) had at least 1 hour per day (or 5 hours per week) of listening experience with synthesized speech for a duration of at least 6 months immediately prior to this study, (b) had professional experience for at least 1 year in the area of assistive technology/AAC, and (c) had experience with both DECTalk and MacinTalk synthesizers in communicative contexts with AAC users.

All listeners were native speakers of American English. In addition, all listeners were currently employed in a setting where they worked exclusively with AAC systems and users. Four were speech and language clinicians, 1 was an occupational therapist, and 1 was a special educator. All 6 listeners in this group were female.

Voices

Six different synthesized voices from two speech synthesizers were employed in this study. The three DECTalk voices were Perfect Paul, Uppity Ursula, and Kit the Kid, for adult male, adult female, and child, respectively. MacinTalk voices were Bruce (MacinTalk II Pro), Agnes (MacinTalk II Pro), and Junior (MacinTalk III Pro), for adult male, adult female, and child, respectively. Voices were selected based on clinical observations of AAC user preferences for quality and intelligibility. Recordings were made with each of the six voices on each of six stimulus word lists.

Stimuli

Selection

Words from the Modified Rhyme Test (MRT) served as the stimulus material (House, Williams, Hecker, & Kryter, 1965). The MRT has been used extensively in testing word level intelligibility of speech synthesizers (Allen, Hunnicutt, & Klatt, 1987; Greene et al., 1986; Logan et al., 1989; Nye & Gaitenby, 1974). It has been called a standard for intelligibility testing in synthesized speech (Duffy & Pisoni, 1992). Because MRT data exist for many different speech synthesizers, performance can be compared directly to previous studies.

The MRT consists of 300 monosyllabic words, mostly CVC, arranged into 50 sets of six rhyming words. Forced choice response and open response formats have been used in speech synthesis research. Intelligibility data obtained from the forced choice paradigm of the MRT reveal scores up to 40% higher than the open response version (Logan et al., 1989).

One important limitation of the MRT is that intelligibility inferences are restricted to perception of monosyllabic CVC words (Logan et al., 1989). However, since limited intelligibility data are available for MacinTalk, it was felt that performance on the MRT would provide a good basis for comparison with other speech synthesizers. The open response version of the MRT was used to provide a more rigorous test of performance. Stimulus words from each of the six lists were randomized for presentation with each of the different voices.

Recording

To control the presentation rate of stimuli, MRT words were recorded on an audio tape with a 7-second interstimulus interval between each word. In order to reduce noise and ensure consistent fidelity between synthesizers and voices, a direct line from the speech synthesizer to the tape recorder was employed, and recording levels were adjusted to maintain a signal-to-noise ratio of 42 dB SPL. Each MRT list was recorded with each voice resulting in six lists per voice.

Presentation

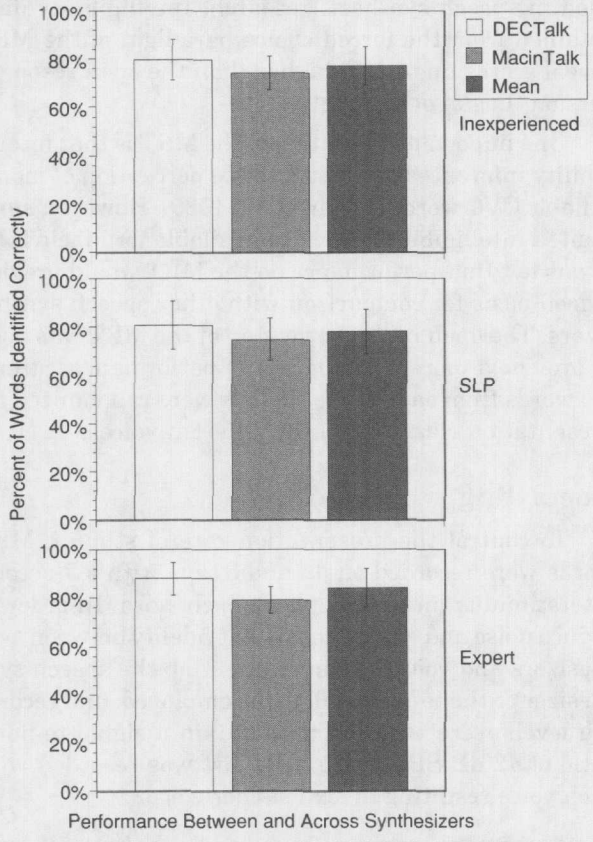
Audio tapes of the MRT lists were presented to each participant in a full-sized soundproof booth. Presentation of the speech material was made through an audio speaker located approximately 38 cm in front of the listener at chest level. A comfortable sound level of approximately 70 dB SPL was used for all listeners. None of the participants reported difficulty hearing the stimuli at the pre-selected sound level.

In order to counterbalance potential learning effects, stimuli were randomized in three ways for each participant: order of synthesized voices, order of word lists, and word list presented by voice. As a result, no two listeners heard the same sequence of lists or the same sequence of voices.

Experimental Protocol/Instruction

Each participant was tested individually to ensure statistical independence. Participants were instructed that they would hear six different lists of words spoken by six different synthesized voices, with each list containing an uninterrupted series of 50 real words. They

Figure 1. Percent of words identified correctly of a possible 50 (±*SD*) for DECTalk, MacinTalk, and the overall mean by the inexperienced group, $n = 6$ (upper panel), the SLP group, $n = 6$ (middle panel), and the expert group, $n = 6$ (lower panel).



were instructed to write down each word immediately after hearing it, using the 7-second interstimulus interval as writing time. Phonetic approximations were encouraged for use in future error analysis in cases where listeners were unsure of what they heard. Participants were told that the purpose of the study was to determine whether or not there were any differences in how well each voice could be understood.

Measurement

Individual words within each list were judged as either incorrect or correct based on whether or not they matched the target word phonemically. A word was counted as incorrect if it contained at least one phonemic misperception when compared with the target word. Percentage of words identified correctly (from a total of 50) are reported.

Design

A 3 × 6 mixed design, or split plot factorial design (Kirk, 1995), using a simple effects model ANOVA was employed for this study. Between-subjects factors were participant groups—Inexperienced, SLP, and Expert. Within-subjects factors were voices—Paul, Ursula, and Kit (DECTalk) and Bruce, Agnes, and Junior (MacinTalk).

Results

Between Subjects

An omnibus F test for the between-subjects factor, group, was significant, $F(2, 15) = 5.856$; $p = .013$. Descriptive results are displayed graphically in Figure 1. ANOVA results in tabular form are presented in Table 2. Post hoc follow up using Tukey's HSD showed that the differences between expert and inexperienced groups, $t(15) = 3.196$; $p < .05$, and expert and SLP groups, $t(15) = 2.656$; $p < .05$, were significant. Table 3 shows these contrasts and statistics in tabular form.

Nested Effects for Voice Within Group

Significant omnibus F tests for voice within each group were as follows: voice within inexperienced group, $F(5, 75) = 8.207$, $p < .01$; voice within SLP group, $F(5, 75) = 11.029$, $p < .01$; and voice within expert group, $F(5, 75) = 8.229$, $p < .01$. Post hoc follow up utilized a series of t tests with error rate controlled using the Dunn-Bonferroni

Table 2. ANOVA Omnibus test results for within- and between-subjects factors.

Source	Sums of Squares	df	Mean Square	F
Groups	40.781	2	20.390	5.856*
Error	52.231	15	3.482	
Voice in group	1011.805	15	87.454	
voice in inexperienced	302.333	5	60.467	8.207**
voice in SLP	406.333	5	81.267	11.029**
voice in expert	303.139	5	60.628	8.229**
Error	552.611	75	7.368	

*statistical significance at $p < .05$
**statistical significance at $p < .01$

Table 3. Tukey post hoc statistics for groups.

Contrast	Mean difference	df	Mean Square for contrast	t
Expert - Inexperienced	3.445	15	35.567	3.196*
Expert - SLP	2.861	15	24.556	2.656*
SLP - Inexperienced	.583	15	1.019	.541

*statistical significance at $p < .05$

procedure (repeated measures nested effects for voice within group, $C = 12$). Contrasts and statistics for follow-up testing are shown in Table 4.

Synthesizer Differences Within Groups

Within each listener group, the difference between DECTalk and MacinTalk was significant with DECTalk being more intelligible than MacinTalk. Results for these contrasts were as follows: inexperienced group, $t(75) = 3.835$; $p < .001$); SLP group, $t(75) = 8.122$, $p < .001$; and expert group, $t(75) = 11.882$, $p < .001$.

Voice Differences Within Groups

Within each listener group a series of three age/gender-matched voice comparisons were made. The difference between Paul and Bruce was the only significant one for each of the three groups with Paul being more intelligible than Bruce. Results for these contrasts were as follows: inexperienced group, $t(75) = 8.422$, $p < .001$; SLP group, $t(75) = 8.122$, $p < .001$; and expert group, $t(75) = 8.573$, $p < .001$. Pairwise differences between female voices and child voices were nonsignificant within each group. Statistics for all follow-up contrasts are presented in Table 4. Descriptive results for individual voices within each group are displayed graphically in Figure 2.

Discussion

The Effects of Experience

For the purposes of this study, expert listeners were defined as individuals who had had extensive exposure to synthesized speech over time, whereas SLPs and inexperienced listeners had no more than incidental experience with synthesized speech. Group differences in this study revealed that expert listeners had higher intelligibility scores, regardless of synthesizer, than speech-language pathologists and inexperienced listeners. There was no difference in the performance of SLPs and inexperienced listeners. This is consistent with the results of Schwab et al. (1983) who found that listeners receiving training with synthesized speech performed better than both listeners receiving training with natural speech and listeners who didn't receive training.

Several conclusions can be drawn from these findings. In the case of SLPs, having advanced knowledge in the area of speech acoustics as well as experience with evaluating and treating deficits in speech, language, and the communication process does not necessarily enhance perception of synthesized speech. That the expert listeners performed better than the other two groups reinforces the findings of Schwab et al. (1983) which suggest that one way to increase proficiency with synthesized speech is to spend time listening to it. Synthetic speech lacks the redundancy of natural speech and is therefore a less effective signal for communication. Auditory experience with this kind of signal is beneficial, probably because it enables listeners to adopt optimal strategies. Being practiced listeners of human speech, as SLPs are, cannot match the benefits of substantial listening experience with synthesized speech.

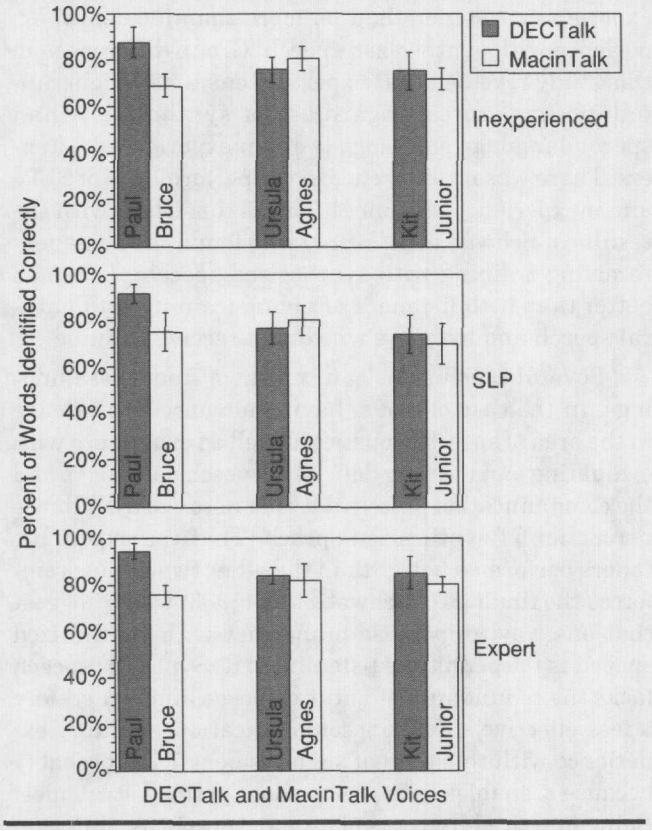
It is important to consider that typical listeners in

Table 4. Statistical tests for $C = 12$ nested post hoc contrasts using the Dunn-Bonferroni procedure.

Contrast	Mean difference	df	Mean Square for contrast	t
Dectalk - Macintalk (inexperienced)	2.283	75	108.375	3.835*
Paul - Bruce (inexperienced)	9.333	75	522.667	8.422*
Ursula - Agnes (inexperienced)	−2.500	75	37.500	2.256
Kit - Junior (inexperienced)	1.667	75	16.667	1.504
Dectalk - Macintalk (SLP)	3.000	75	486.000	8.122*
Paul - Bruce (SLP)	8.167	75	400.167	7.369*
Ursula - Agnes (SLP)	−1.167	75	8.167	1.053
Kit - Junior (SLP)	2.000	75	24.000	1.805
Dectalk - Macintalk (expert)	4.389	75	1040.167	11.882*
Paul - Bruce (expert)	9.500	75	541.500	8.573*
Ursula - Agnes (expert)	1.167	75	8.167	1.053
Kit - Junior (expert)	2.500	75	37.500	2.256

*statistical significance at $p < .001$

Figure 2. Percent of words identified correctly of a possible 50 (±*SD*) by the inexperienced group, $n = 6$ (upper panel), the SLP group, $n = 6$ (middle panel), and the expert group, $n = 6$ (lower panel) for individual voices within DECTalk and MacinTalk synthesizers. Voices within DECTalk are shown paired with gender/age-matched voice within MacinTalk.



the public arena are most like the inexperienced listener group in this study. Although AAC specialists perceive synthesized speech as more intelligible than do speech-language pathologists and inexperienced listeners, the same intelligibility pattern was observed across experience groups. That is, the differences between experience groups are ones of magnitude, not of profile. This suggests that experience increases intelligibility but does not alter perceptual characteristics of individual voices.

Synthesizer Differences

Findings from this study reveal that the DECTalk synthesizer is more intelligible than the MacinTalk synthesizer within each group of listeners—inexperienced, SLP, and expert. Data from separate DECTalk voices show that the mean percentage of words identified correctly for the male voice Paul is markedly higher than for the other two voices. Because of Paul's superior intelligibility over the other DECTalk voices, the mean for DECTalk is significantly higher than the mean for MacinTalk. As a result, there is not a collective advantage for DECTalk compared to MacinTalk. Rather, the voice Paul is so much more intelligible than the other DECTalk voices that the mean for DECTalk is elevated because of this one voice only and not because DECTalk intelligibility for each voice is superior to MacinTalk. Examination of descriptive data from separate MacinTalk voices shows that intelligibility scores for the male voice Bruce are close to or worse than the scores for the other MacinTalk voices. Consequently, the mean for MacinTalk is more representative of the synthesizer as a whole than a reflection of one outstanding voice.

Intelligibility profiles for individual voices within each synthesizer may reflect differences in the synthesis algorithms. The female and child voices in DECTalk were derived largely by changing parameters of the male voice, Perfect Paul. It is well known that speech characteristics of women and children are different from those of men, and are not just feminized or juvenilized versions of the adult male voice (see Klatt & Klatt, 1990, for a discussion of women's voices). The decreased intelligibility from Paul to Ursula and Kit appears to reflect this. However, in diphone-based MacinTalk, this discrepancy in intelligibility scores is not present. Because female and child voices are not adaptations of the male voice in diphone-based synthesis, intelligibility scores do not favor the male voice. In fact, examination of descriptive data indicates that the adult female voice, Agnes, had better intelligibility than the male voice, Bruce.

Voice Differences

Findings from the present study reveal that the DECTalk male voice, Paul, was significantly more intelligible than the MacinTalk II Pro voice, Bruce, within each of the three listener groups. There were no other pair-wise differences between MacinTalk and DECTalk voices. This discrepancy between the findings of Rupprecht et al. (1995) and the present study is likely due to the differing amount of linguistic context provided by the stimulus material in the two experimental tasks. It is important to note that Rupprecht et al. examined word level intelligibility using the SPIN test, a task quite different from isolated word-level transcription in the MRT. Word responses from the high predictability SPIN sentences (half of the stimuli in the Rupprecht et al. study) allow the listener to use syntactic and semantic information to enhance understanding of words, resulting in increased intelligibility scores. An isolated word task does not. This has been shown for both natural speech (Miller, Heise, & Lichten, 1951) and synthetic speech (Greene, Manous, & Pisoni, 1984; Hoover, Reichle, VanTassell, & Cole, 1987).

It is also noteworthy that different versions of the MacinTalk Pro synthesizer were used between the present study and the Rupprecht et al. (1995) study. The

MacinTalk Pro synthesizer used by Rupprecht et al. was the predecessor to the MacinTalk II Pro synthesizer used in the present study. Logically, the more recent version of the synthesizer should reflect advances in synthesis software development as well as increased microprocessor power. Although the more recent software version, MacinTalk II Pro, would be expected to be better than the older version, MacinTalk Pro, this is difficult to judge because of experimental task differences.

Taken together, these two studies suggest the DECTalk male voice, Paul, performs better than the MacinTalk male voice, Bruce, in the absence of linguistic context. However, when linguistic information is added to the intelligibility task, these two voices may become more comparable, assuming that MacinTalk II Pro voices are equivalent or better than MacinTalk Pro voices.

Intelligibility for DECTalk Voices

Results of the present study are consistent with others examining isolated word-level intelligibility with DECTalk voices. The DECTalk male voice, Paul, has been found to be 88% intelligible with inexperienced listeners in several different experiments using the open format MRT (Greene et al., 1986; Logan et al., 1989), including the present study.

Intelligibility for the DECTalk female voice, Ursula, has not been examined in other studies. However, intelligibility for the DECTalk female voice, Betty, has been found to be 81% with inexperienced listeners on the open format MRT (Greene et al., 1986; Logan et al., 1989). Intelligibility for Ursula in this study was found to be 76% for inexperienced listeners. Again, these results are reasonably consistent and may not reflect an important intelligibility difference between the two different DECTalk female voices.

Few studies have examined intelligibility for the DECTalk child voice, Kit. In fact, no previous studies have examined this voice using the MRT. Findings using other stimulus material have varied. Mirenda & Beukelman (1987) found that Kit had 68% intelligibility for inexperienced adult listeners on words from the Computerized Assessment of Dysarthric Speech (Yorkston, Beukelman, & Traynor, 1984) whereas McNaughton et al. (1994) found 80% accuracy on open-format word level stimuli taken from preschool vocabulary lists. In the present study, inexperienced listeners had 75% intelligibility for Kit. These results are difficult to compare given that stimulus materials differ; however, findings from the present study are within the range of intelligibility scores observed in previous studies.

Intelligibility for MacinTalk Voices

Rupprecht et al. (1995) found that the MacinTalk Pro male voice had 87.3% intelligibility, and the MacinTalk Pro female voice had 85.4% on the SPIN test (Kalikow et al., 1977). Results of the present study reveal 69% accuracy for MacinTalk II Pro Bruce, and 81% intelligibility for MacinTalk II Pro Agnes on isolated word-level word level transcription using the MRT.

Conclusions

This study has examined word level intelligibility in a quiet environment for DECTalk and MacinTalk speech synthesizers. Results of this study suggest that MacinTalk and DECTalk compare closely, with the exception of the adult male voices, which were the only voices that differed in intelligibility. Female and child voices had comparable intelligibility. This pattern was consistent for listeners from different experience groups as well. With some improvement in the adult male voice, it appears that MacinTalk's diphone-based synthesis has the potential to rival DECTalk's formant-based synthesis. This has important implications for AAC users in terms of portability, cost, and the ability to upgrade speech synthesis software as improvements become available. This study has examined only intelligibility and not quality, user preference, or naturalness. The ultimate comparison between synthesis systems should address these issues as well.

Clearly, it is difficult to generalize from word level intelligibility findings to synthesized speech in general. In AAC, there is nearly always linguistic context surrounding the use of synthesized speech. However, single word responses do occur in conversational settings, for example, as greetings, replies to questions, or as clarifications. Use of isolated word-level intelligibility findings, although less generalizable to natural contexts, provides important information at the most basic level of analysis, the word and phoneme. Findings from this type of study probably reflect the worst case scenario for intelligibility. It is generally very difficult to obtain information on segmental and feature errors from intelligibility tests that use phrasal or sentential materials. Single words are more suitable for this purpose because they permit comparisons of different phonetic elements in constrained syllable positions. Such item analyses are being examined as an additional part of this project. One potential benefit of segmental-level error analyses is that they can point to the need for improved synthesis at the phonetic level. It is likely that future speech synthesis systems will offer customized voices. Progress in speech synthesis should shorten the time needed for this purpose.

Future Research

Additional research should examine perceptual differences at the segmental level between these two synthesizers. Analyses should include confusion matrices

detailing listener agreement as well as error patterns for phonemes in all positions of words across the various voices in each synthesizer. Understanding the consistency and nature of error patterns among listeners and voices would suggest specific phonetic level improvements for synthesis developers. In addition, comparisons between expert listeners and typical listeners on segmental errors would help clarify the impact of experience at a finer level than word intelligibility alone.

In the future, comparison between DECTalk and MacinTalk synthesizers should occur in situations that approximate real communication. These include intelligibility at the message level, in communicative contexts, and in noisy environments.

Acknowledgment

This research was supported in part by NIH research grant DC00319 ("Intelligibility Studies of Dysarthria") awarded to the second author from the National Institute on Deafness and Other Communicative Disorders.

References

Allen, J., Hunnicutt, S., & Klatt, D. (1987). *From text to speech: The MITalk system*. Cambridge, England: Cambridge University Press.

Duffy, S. A., & Pisoni, D. B. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Language and Speech, 35*, 351–389.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

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.

Greene, B. G., Manous, L. M., & Pisoni, D. B. (1984). *Preliminary evaluation of DECtalk (Speech Research Lab Technical Note 84-03)*. Bloomington, IN: Indiana University.

Hoover, J., Reichle, J., Van Tassell, D., & Cole, D. (1987). The intelligibility of synthetic speech: Echo II vs. Votrax. *Journal of Speech and Hearing Research, 30*, 425–431.

House, A. S., Williams, C. E., Hecker, M. H., & Kryter, K. E. (1965). Articulation testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America, 37*, 158–166.

Kalikow, D. N., Stevens, K. N., & Elliot, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America, 61*, 1337–1351.

Kirk, R. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing.

Klatt, D. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America, 67*, 979–995.

Klatt, D., & Klatt, L. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America, 87*, 820–857.

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.

McNaughton, D., Fallon, K., Tod, J., Weiner, F., & Neisworth, J. (1994). Effect of repeated listening experience on the intelligibility of synthesized speech. *Augmentative and Alternative Communication, 10*, 161–168.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology, 41*, 329–225.

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.

Nye, P., & Gaitenby, J. (1974). The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. *Haskins Laboratories Status Report on Speech Research, SR-38*, 169–190.

O'Shaughnessy, D., Barbeau, L., Bernardi, D., & Archambault, D. (1988). Diphone speech synthesis. *Speech Communication, 7*, 55–65.

Rupprecht, S., Beukelman, D., & Vrtiska, H. (1995). Comparative intelligibility of five synthesized voices. *Augmentative and Alternative Communication, 11*, 244–247.

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

Schwab, E., Nusbaum, H., & Pisoni, D. (1983). *Some effects of training on the perception of synthetic speech. (Research on speech perception, Progress Report No. 9)*. Bloomington, IN: Indiana University.

Venkatagiri, H. (1994). Effect of sentence length and exposure on the intelligibility of synthesized speech. *Augmentative and Alternative Communication, 10*, 96–104.

Venkatagiri, H. (1996). The quality of digitized and synthesized speech: What clinicians should know. *American Journal of Speech-Language Pathology, 5*, 31–42.

Yorkston, K., Beukelman, D., & Traynor, C. (1984). Computerized assessment of intelligibility of dysarthric speech [Computer software]. Austin, TX: Pro-Ed.

Received March 31, 1997

Accepted February 13, 1998

Contact author: Katherine C. Hustad, Department of Special Education and Communicative Disorders, University of Nebraska-Lincoln, 253 Barkley Memorial Center, Lincoln, NE 68583-0731. Email: khustad@unlinfo.unl.edu