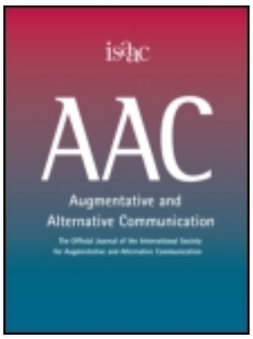

Augmentative and Alternative Communication

ISSN: 0743-4618 (Print) 1477-3848 (Online) Journal homepage: https://www.tandfonline.com/loi/iaac20

Factors influencing ratings of speech naturalness in augmentative and alternative communication

Ann Ratcliff, Sue Coughlin & Mark Lehman

To cite this article: Ann Ratcliff, Sue Coughlin & Mark Lehman (2002) Factors influencing ratings of speech naturalness in augmentative and alternative communication, Augmentative and Alternative Communication, 18:1, 11-19, DOI: 10.1080/aac.18.1.11.19

To link to this article: https://doi.org/10.1080/aac.18.1.11.19

Published online: 12 Jul 2009.

0743-4618/01/1801-0011 $3.00/0; Volume 18, March 2002
AAC Augmentative and Alternative Communication

Factors Influencing Ratings of Speech Naturalness in Augmentative and Alternative Communication

Ann Ratcliff, Sue Coughlin, and Mark Lehman
Department of Communication Disorders, Central Michigan University, Mount Pleasant, Michigan, USA

The concept of speech naturalness has been used in the field of speech-language pathology as a clinical measure of perceptual quality of "normal" and "not normal" speech. Whereas measures of intelligibility have been commonly used to assess the quality of voice output augmentative and alternative communication (AAC) devices using DECTalk™ speech, measures of speech naturalness have not. Three studies were conducted to determine the effects of manipulation of rate, pitch, and pause on ratings of speech naturalness by naive listeners of DECTalk synthetic speech. The results indicate that DECTalk speech characterized by faster rate and no added pauses was perceived as being more natural than speech with slow rate and added pauses. Manipulation of pitch had no effect on naturalness ratings.

KEY WORDS: augmentative and alternative communication (AAC), speech naturalness, synthesized speech

The parameter known as *speech naturalness* has been used in the field of communication disorders as a useful clinical measure. Speech naturalness has been defined as speech output that sounds normal or natural to the listener (Parrish, 1951). The essential component of Parrish's concept of naturalness is that speech naturalness allows the listener to focus attention on the meaning of words spoken rather than on the speech pattern used to convey the meaning. In other research, Sanders, Gramlich, and Levine (1981) defined speech naturalness as speech produced by speakers using the normal and customary speech patterns accepted by the community. Moreover, for Sanders et al., naturalness is one of three components of speech quality, which also includes intelligibility and clarity. In this model, intelligibility of speech is defined simply as speech that can be understood, and clarity is defined as speech that can be understood without difficulty. The value of this model is that naturalness, although not completely separate and distinct, is conceptualized apart from intelligibility.

Along these lines, Coughlin (1998) studied naturalness characteristics of the speech produced by children and adolescents. Naive listeners were asked to rate the speech naturalness of audiovisual speech samples of normal and disordered speakers. Naturalness was not defined for the listeners. When asked to comment on cues of naturalness, listeners referred to the importance of intelligibility as the foremost quality of the speech. Although intelligibility and naturalness are intimately related, it appeared from this study that speech must be intelligible before its naturalness can be rated.

The concept of speech naturalness apart from intelligibility can best be illustrated in the area of fluency disorders. Many current therapies for individuals who demonstrate nonfluent speech employ fluency shaping procedures that have been criticized for producing a speech quality that sounds unnatural or different from the norm (Ingham & Packman, 1978; Runyan & Adams, 1978, 1979; Runyan, Bell, & Prosek, 1990). Thus, speakers undergoing fluency shaping procedures may be intelligible while still producing speech that calls attention to itself as not sounding natural.

Although the speech naturalness parameter has proven a useful clinical measure in the area of fluency, it has not been extensively applied as a variable related to the voice output communication aids (VOCAs) used in the area of augmentative and alternative communication (AAC). Practitioners and consumers in the area of AAC have acknowledged the advantages of voice output in AAC devices (McNaughton, Fallon, Tod, Weiner, & Neisworth, 1994; Raghavendra & Allen, 1993) since such output enables AAC users to communicate from a distance, on the telephone, in a car, or with people who are visually impaired. This is possible because VOCA output often simulates normalizing, effective, efficient oral speech (Higginbotham, Drazek, Kowarsky, Scally, & Segal, 1994; Raghavendra & Allen, 1993).

Speech naturalness ratings of the synthetic speech in VOCAs might be expected to yield information that

may facilitate more successful use of voice output in AAC devices. In reviewing the AAC literature, a number of patterns that typify the extant research in this area can be identified. First, several studies have evaluated the characteristics of listeners in assessing the adequacy of VOCAs. For example, Mirenda and Beukelman (1990) found that natural speech was rated as being more intelligible than synthesized speech in single words and in sentences by children as young as 7 years of age and adults. Similarly, Crabtree, Mirenda, and Beukelman (1990) found that listeners, regardless of age or gender, preferred natural speech to the output of any of eight voices from six different speech synthesizers. The results of these studies revealed that the preference for voices that sound close to human production is independent of listener demographics. In fact, Quist and Blischak (1992), in outlining specifications for VOCAs, stated that an important characteristic of an effectively used communication aid is that the voice quality should be as pleasing and as human sounding as possible.

A second pattern noted in the AAC literature involves comparisons among various voices. One comparison is between naturally and synthetically produced speech, which was examined in many of the studies described previously. The consistent result of these studies has been that naturally produced speech is greatly preferred over synthetic speech (e.g., Crabtree et al., 1990; Mirenda & Beukelman, 1990). Other comparisons are between multiple VOCAs or between the same VOCA following manipulation of a variety of speech parameters. For example, Venkatagiri (1991) examined the effect of manipulating speech presentation rate on VOCA intelligibility. He found that reducing the rate of speech from 201 to 139 syllables per minute improved word intelligibility by more than 10%, whereas manipulations of pitch had no significant effect. In another study, Higginbotham et al. (1994) examined the effect of rate of presentation on listener comprehension of synthetic speech, using discourse summary measures as dependent variables. Both DECTalk™ and an Echo speech synthesizer presented the paragraph-level discourse at two rates. One rate was set at 128 to 149 words per minute, which is considered within the norm for human speaking rate. The other was set with a 10-second delay between words to simulate the slow output rate of many VOCAs, resulting in a rate of 5.6 to 5.8 words per minute. The results indicated that the slower rate significantly improved the listeners' ability to comprehend the texts. The authors speculated that the silent intervals enabled listeners to better process the speech signal. The results of both the Venkatagiri (1991) and Higginbotham et al. (1994) studies indicated that a slower speech presentation rate resulted in improved ratings of synthesized speech intelligibility.

Venkatagiri (1994) also studied the relationship between the intelligibility of VOCAs, sentence length, and listeners' exposure to synthetic speech. The results indicated that the more sentences the listener was exposed to, the more intelligibly the synthesized speech was rated. Sentence length itself was not significant, however. Hustad, Kent, and Beukelman (1998), on the other hand, examined the single-word intelligibility of DECTalk and MacinTalk™ speech synthesizers. They found that the two synthesizers had very similar intelligibility ratings, as measured by subjects' written responses to a randomized list of 50 words. Reynolds and Jefferson (1999) found that when children were asked to judge the truth of statements delivered by DECTalk and human speech, response latency times were significantly longer for the synthetic speech, even after practice. The authors speculated that listeners may use more information processing resources to understand synthesized speech. McNaughton et al. (1994), who found that DECTalk speech was superior to Echo speech on all measures, also reinforced the concept that the closer synthesized speech approximates humanly produced speech, the more communication is enhanced.

A final pattern noted in the AAC literature is the reliance on measures of intelligibility to evaluate listener preferences for synthesized speech. Although intelligibility was a reasonable measure when comparisons involved early and later generations of VOCAs, as the quality of synthesized voice continues to improve to the point at which intelligibility is no longer an issue, other measures to evaluate these voices need to be found. Moreover, reliance on increased intelligibility as the most important measure can lead to questionable conclusions. For example, Higginbotham et al. (1994) determined that presenting synthesized speech at a speech rate of approximately six words per minute facilitated measures of speech intelligibility. However, one might ask how long the interaction between a typical listener and a VOCA user would last in a situation in which the speech was being produced so slowly. Clearly, parameters other than intelligibility need to be used. For example, Hustad et al. (1998) commented that intelligibility is just one of a number of parameters that may influence how a person rates speech. Other studies have included speech quality, user preferences, and naturalness. Gorenflo, Gorenflo, and Santer (1994) found that ease of listening was a significant covariate when examining the attitudes of men and women toward a videotaped AAC user using synthesized speech. In their study, ease of listening appeared to be more important than even gender appropriateness.

This article discusses investigations of the concept of speech naturalness as a factor in synthetic speech acceptability beyond the concept of intelligibility. The concept of speech naturalness as applied to VOCAs is an important issue for several reasons. First, the population of potential AAC users is growing because of a "zero reject" model, so that more people are using VOCAs than a decade ago (Kangas & Lloyd, 1988;

Reichle, York, Sigafoos, 1994). Second, some research literature suggests that approximately 50% of speech-language pathologists have at least one AAC user in their caseloads, resulting in reduced use of an expert model for AAC treatment as users of AAC are spread throughout the caseloads of many practitioners (Ratcliff & Beukelman, 1995; Simpson, Beukelman, & Bird, 1998). Third, since technology is now able to produce highly intelligible synthesized speech used in many VOCAs, more attention can be directed to other aspects of speech, such as naturalness. Most synthesizers used in VOCAs offer a variety of acoustic options that can be manipulated by the consumer and his or her team. These options provide an opportunity for the consumer to make subtle acoustic variations in the VOCA output that may facilitate communication. Indeed, some VOCA users themselves have stated that they take advantage of these acoustic options in their own communication (the Augmentative and Alternative Communication Online Users Group [ACOLUG], personal communication, April 2–3, 2000).

The purpose of this article is to report the results of three experiments that were designed to answer three main questions: (1) What factors of DECTalk synthesized speech contribute to ratings of "highly natural" and "highly unnatural" speech by naive listeners? (2) Do ratings of DECTalk speech output naturalness by naive listeners change when the factors identified by answering the first question are manipulated? and (3) Does the use of listeners who meet stringent reliability criteria influence the naturalness ratings of synthesized speech?

STUDY 1

Given that speech naturalness has been defined previously as speech that sounds normal or natural to the listener, any attempt to rate speech naturalness requires listeners to make some reference to an internal or external standard of human speech. Since, from a research perspective, it was desirable to have listeners making reference to a similar standard of naturalness, speech samples from normal-speaking children in a rating task were used to provide a common referent. The purpose of study 1 was to identify potential parameters in DECTalk synthesized speech that may contribute to ratings of "highly natural" and "highly unnatural" speech by naive listeners by having them rate samples of both DECTalk and human-produced speech. An additional purpose was to verify expected differences in the naturalness ratings of normally produced and synthesized speech.

Method

Stimuli

A cassette tape recording was made using 24 different 30-second speech samples collected from children as part of Coughlin's (1998) study. Twelve of the 30-second samples were of normal-speaking children (six 8 year olds and six 10 year olds), with an equal number of female and male speakers. The other 12 samples were programmed into a DynaMyte™ VOCA (DynaVox, Inc.) using DECTalk Kit the Kid voice. The accessible speech parameters of the DynaMyte were set at the manufacturer's default settings. All 24 samples were recorded onto the cassette in random order.

Listeners

Thirty-five graduate students in speech-language pathology served as listeners. Students were enrolled in a class on fluency disorders and completed the listening tasks for course credit. None of the students reported having frequent experience with VOCAs; only 7 reported occasional or frequent experience with VOCAs, whereas the remaining 28 reported little or no experience with VOCAs. Thus, this population was considered naive with respect to VOCA experience. This was later substantiated by correlating listeners' reported experience with VOCAs and their ratings of naturalness of the DECTalk samples. The Spearman rank order correlation was $r_s = .37$ ($p < .05$), indicating that listeners' ratings of naturalness were only weakly related to their experience with VOCAs.

Procedure

Verbatim instructions to the listeners are provided in the Appendix. Listeners were asked to rate the 24 audio stimuli using a 9-point speech naturalness rating scale. This scale is an equal-appearing interval scale similar to those frequently used in the naturalness literature (Coughlin, 1998; Finn & Ingham, 1994; Ingham, Ingham, Onslow, & Finn, 1989; Ingham, Martin, Haroldson, Onslow, & Leney, 1985; Ingham & Onslow, 1985; Martin, Haroldson, & Triden, 1984; Onslow, Hayes, Hutchins, & Newman, 1992; Sanders et al., 1981). A rating of 9 on the scale indicates a rating of highly unnatural and a rating of 1 is a rating of highly natural. This scale has been shown to be a reliable tool for scaling perceptual differences both within and between a variety of speakers (Coughlin, 1998). All listeners were provided with 24 response sheets, each containing one 9-point equal-appearing interval rating scale. Listeners rated each of the 24 audio stimuli on the 9-point scale by circling their rating and then turning the page to record the next response.

Results and Discussion

The mean naturalness rating for the DECTalk samples was 8.0; the mean rating for the human speech samples was 2.6. The difference was significant, $t\ (22) = 27.3$, $p < .001$, revealing that the synthesized speech stimuli were consistently rated as less natural sounding than the human speech. Moreover, the

mean ratings for the DECTalk voices ranged from 7.5 to 8.4, whereas those for the live voices ranged from 1.7 to 4.1. These results indicate that the listeners used a restricted portion of the naturalness rating scale for the DECTalk voices. This may have been attributable to the fact that the human voices represented 12 different speakers, whereas the DECTalk samples were essentially the same speaker saying 12 different things. Moreover, the inclusion of both DECTalk and human voices in the listening experience may have created an anchoring effect that limited the listener ratings of the DECTalk voices to the extreme "unnatural" end of the rating scale. This suggests that, to obtain more robust naturalness ratings of DECTalk voices, listeners may have to be limited to only synthesized voices when making naturalness judgments.

The listeners were also asked to write their reasons for rating the speech samples as they did, at the completion of the experiment. The responses were analyzed by grouping them into categories. The categories with the greatest number of comments had to do with listeners' perception of variations in rate, pitch, and pause.[1]

STUDY 2

Listeners in study 1 rated the DECTalk speech as significantly more unnatural than the human speech. They also indicated that rate, pitch, and pause time were the variables perceived as most affecting naturalness. Given these results, the purpose of study 2 was to determine how listeners would rate the naturalness of samples of synthesized speech that differed in rate, pitch, and pause time.

Method

Stimuli

Three transcripts from study 1 that were judged to be most natural were chosen to form the basis for the stimuli used for this study. These transcripts were entered into a DynaMyte for manipulation and subsequent voice output. Manipulation consisted of varying the rate and pitch characteristics of the device using the built-in controls. Three levels of rate (150, 180, and 210 syllables/minute) were chosen based on Guitar's (1998) normative data for the speech rates of young children. Furthermore, three levels of pitch (216, 261, and 306 Hz) were created based on Baken's (1987) data on the pitch ranges for children. Acoustic analysis of the DynaMyte output indicated actual rates of 147, 176, and 204 syllables per minute and actual fundamental frequency levels of 208, 253, and 305 Hz. These rate and pitch manipulations resulted in a total of nine variations for each of the three transcripts. These stimuli were outputted from the device and digitized into a Windows-PC for further manipulation and playback. Wave Studio™ software, Version 3.19 (Creative Technology Ltd., 1992–96), was used to digitize the samples at a sampling rate of 22 kHz.

The final manipulation of the stimuli involved the insertion of pauses at specified points in each stimulus. These points were determined by acoustic analysis of the original (human voice) stimuli using MultiSpeech™ Model 3700 software (Kay Elemetrics, 1996) on a Windows-PC. This analysis determined the length of each of the naturally occurring pauses. Although the DECTalk software used in the DynaMyte inserted pauses at commas and periods in the transcript, each original stimulus had pauses at several other locations. These pauses were recreated in the synthesized speech version by inserting silent intervals at the appropriate points in each stimulus. The no-pause condition was the default pausing delivered by the DECTalk speech, which formulates some pause time relative to punctuation. No other researcher-imposed pauses were added to speech output during the no-pause condition. This resulted in the creation of a total of 54 stimuli (3 transcripts $\times$ 3 pitch levels $\times$ 3 rate levels $\times$ 2 pause levels).

Listeners

Seventeen listeners, all undergraduate students majoring in special education, volunteered for the study. Like the listeners for study 1, they were all considered naive with respect to VOCA experience. The Spearman rank order correlation between listeners' reported experience with VOCAs and their ratings of naturalness was quite low, $r_s = .11$ ($p > .05$), which indicates that their ratings of naturalness were not related to their experience with VOCAs.

Procedure

The stimuli were presented to listeners in groups. Each group completed the task in one session lasting approximately 50 minutes. All listeners were provided with 54 response sheets, each containing one 9-point equal-appearing interval rating scale. The listeners rated each of the 54 stimuli on the 9-point naturalness scale used in study 1 by circling their rating and then turning the page to record the next response. The instructions were the same as in study 1, with the exception that the listeners were advised beforehand that the speech samples they were about to hear were produced by the synthesized speech used in AAC devices. Finally, the listeners were provided with a hard copy of each transcript to control for intelligibility. Intelligibility was treated as a control and not as a dependent variable in this study because the major

[1]Verbatim transcripts of the comments are available from the first author.

concern was the manipulation of speech parameters and how that would affect listeners' perception of the naturalness of the speech output. Since reduced intelligibility could influence listeners' rating of naturalness, the transcripts were provided so that the listeners could focus on how the message sounded without concern for the message content.

Mean naturalness ratings were computed across the 17 listeners for each of the 54 stimuli. These values were subjected to a three-way (Pitch × Rate × Pause Time) analysis of variance (ANOVA). Transcript was not treated as an independent variable since preliminary analysis indicated no differences in naturalness ratings across the three transcripts.

Results and Discussion

Figure 1 shows the mean naturalness ratings for each of the three pitch levels. Mean naturalness ratings varied from 6.6 to 6.7. The ANOVA revealed that pitch level had no impact on listeners' ratings of naturalness for the stimuli, $F(2, 36) = .051$, $p > .05$. This is similar to the findings of Venkatagiri (1991), who studied the issue of intelligibility by having listeners rate randomized short sentences. His results indicated that pitch did not appear to consistently influence ratings of intelligibility.

Mean naturalness ratings as a function of stimulus speaking rate are shown in Figure 2. Although the mean naturalness ratings varied between 6.3 and 7.1, the ANOVA indicated that these differences were significant, $F(2, 36) = 5.967$, $p < .05$. Post hoc pairwise Scheffé comparisons revealed that naturalness ratings for the slowest rate condition (7.1) were significantly different from those for the fastest rate condition (6.3), with listeners perceiving the fastest rate condition as more natural. Venkatagiri (1991) found that decreasing the rate of syllables per minute improved word intelligibility by 10%. However, no information regarding the impact of decreasing speaking rate on speech naturalness was provided in his report.

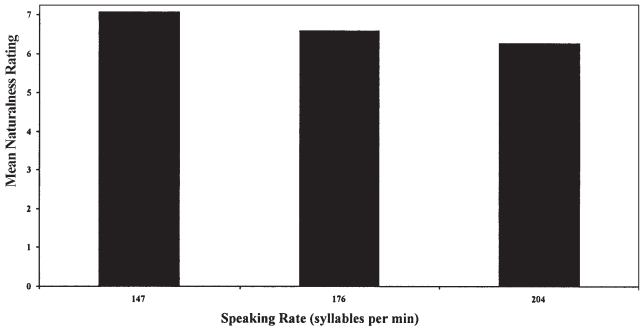

Figure 2. Mean naturalness ratings for stimuli at each of three speaking rate levels in study 2.

Mean naturalness ratings as a function of pause time are shown in Figure 3. The mean naturalness rating for the no-pause condition was 5.9 and the mean rating for the pause condition was 7.4. Manipulations of pause and no-pause conditions were related to naturalness judgments and were significant, $F(1, 36) = 57.04$, $p < .05$. The no-pause condition was the DECTalk default setting, which occurs without specific manipulation of this parameter. Listeners rated the DECTalk default setting as more natural sounding than when pauses corresponding to those found in 8- and 10-year-old speakers were inserted.

Although both the manipulation of rate and pause significantly affected listener ratings of naturalness in this study, the magnitude of the differences (0.8 for rate and 1.5 for pause) represented only a small difference on the naturalness scale. Although it may be that the differences in naturalness for these stimuli are indeed small, it may also be that another factor contributed to the results obtained in this study. During data reduction, it was observed that the listeners varied in their use of the rating scale. Some listeners used a broad range of the scale, whereas others used

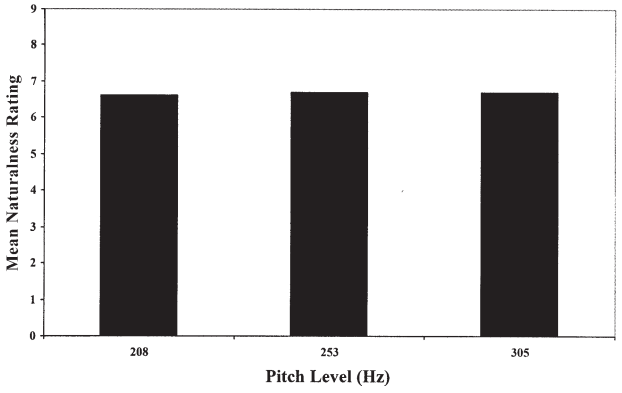

Figure 1. Mean naturalness ratings for stimuli at each of three pitch levels in study 2.

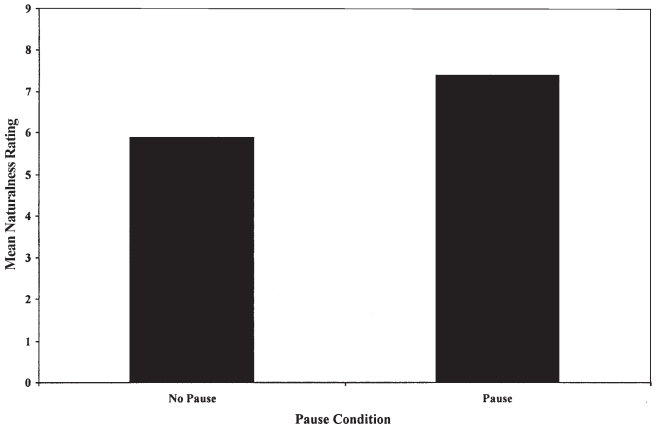

Figure 3. Mean naturalness ratings for stimuli in the no-pause and pause conditions in study 2.

a narrower range. Some listeners used the middle to upper end of the scale, whereas others used the middle to lower end. Moreover, the reliability of the listeners was not assessed, so the consistency with which the listeners used the scale is unknown. Even given the within-subjects nature of this design, including listeners who applied the scale in different ways may have contributed to the small differences obtained.

STUDY 3

The results of study 2 appeared to indicate that stimuli with increased rate and fewer and shorter pauses were perceived as more natural than stimuli with slow rate and more and larger pauses, whereas average pitch level did not appear to impact on ratings of naturalness. However, the differences in the average ratings on the naturalness scale were quite small, perhaps because of variable listener use of the scale. Therefore, the purpose of the third study was to extend the results of study 2 by using a larger number of stimuli manipulating rate and pause and by using a listener group chosen for its high degree of inter- and intrarater reliability.

Method

Stimuli

The stimuli for study 3 used only one level of pitch (261 Hz), the default for the Kit the Kid voice of the DynaMyte, and manipulated rate and pause in the same way as in study 2. Eight transcripts from study 1 that were judged to be most natural, including the three transcripts used in study 2, were chosen to form the basis of the stimuli used in this study. This resulted in a total of 48 unique stimuli (8 transcripts $\times$ 3 rate levels $\times$ 2 pause levels) that were presented for listener ratings for the final study. To control for the impact of intelligibility on naturalness ratings, listeners were provided with the written text of what they heard, as in study 2.

Listeners

The listeners in study 3 were drawn from a pool of 65 college-age adults who were enrolled in undergraduate courses and completed the listening tasks for course credit. These listeners were evaluated in terms of their intra- and inter-rater agreement to identify a homogeneous group in this regard.

Intrarater reliability was determined by repeating 10 randomly selected stimuli for listener rating during the listening task. The percentage of repeated ratings that were within ±1 rating point of the original ratings was computed. Listeners were considered reliable if 80% of their repeated ratings met this criterion. Twenty-eight of the 65 listeners achieved this level and remained in the pool for further analyses.

Inter-rater agreement was determined for the remaining 28 listeners by calculating the percentage of time that each listener's rating agreed within ±1 rating point of the other 27 listeners for all 48 stimuli. For a 9-point scale of the type used in this study, the percentage of agreement attributable to chance alone would be 30.9% (Kreiman, Gerratt, Kempster, Erman, & Berke, 1993). Moreover, since in practice many listeners avoid extreme ratings (Kreiman et al., 1993), viewing the scale as consisting of 7 points results in a percentage of agreement attributable to chance of 38.8%. Therefore, an arbitrary agreement level of 50% was set as the criterion for acceptable inter-rater reliability. Eleven listeners failed to reach that level and were eliminated from further analyses. This resulted in a total of 17 listeners who were determined to have both acceptable intrarater reliability and inter-rater agreement and were thus used in study 3.

Procedure

The stimuli were presented to listeners in groups. A total of three groups met to complete the project. Each group completed the task in one session lasting approximately 50 minutes. All listeners were provided with 58 response sheets, each containing one 9-point equal-appearing interval rating scale (48 unique stimuli and 10 repeated for reliability). All other procedures were the same as in study 2.

Results and Discussion

The mean ratings of the 17 listeners for each of the 48 stimuli were analyzed using a 2 $\times$ 3 ANOVA. Table 1 shows the mean ratings for the no-pause/pause and three rate conditions.

The ANOVA revealed a significant main effect for Pause, $F(1, 42) = 28.66$, $p < .001$, with the stimuli containing no pauses rated as being more natural than those with pauses. There was also a significant main effect for Rate, $F(2, 42) = 11.86$, $p < .001$. Multiple post hoc Scheffé comparisons indicated that the

TABLE 1: **Mean Ratings of Naturalness for the Rate and Pause Conditions**

	Pause Condition		
Rate (wpm)	*No Pause*	*Pause*	*Total*
147	6.3	6.8	6.5
176	5.3	6.3	5.8
204	4.8	6.4	5.6
Total	5.4	6.5	

wpm = words per minute.

slow rate condition was rated as significantly less natural than the medium and fast rate conditions but that those conditions were not different from each other. The interaction between the pause and rate conditions failed to reach significance, $F(2, 42) = 3.00$, $p = .061$.

The results of study 3 were consistent with study 2 in that manipulations of rate and pause condition were perceived as being significantly different by a pool of listeners judged to have good intra- and inter-rater agreement. However, Table 1 shows that restricting the analysis to listeners with high intra- and inter-rater reliability did little to affect the overall differences in ratings for either of these two variables. Differences across the levels of rate in study 3 spanned 1.0 scale point, whereas the differences in study 2 spanned 0.8 scale points. The difference in listener ratings between the two pause conditions in study 3 (1.0 scale point) was actually less than was observed in study 2 (1.5 scale points). Therefore, it would appear that restricting the analysis to a group of supposed "superior listeners" had no impact on the magnitude of the differences observed in study 3.

To further investigate the relationship of both rate and pause to naturalness ratings, a Spearman rank order correlation was computed between the naturalness ratings and speech rate. A syllables-per-minute measure of speech rate was calculated for each stimulus by measuring the total time in minutes of the stimulus including pause and dividing that time into the number of syllables. The correlation between this measure and the mean naturalness rating was $r_s = -.794$, which was significant ($p < .01$). Moreover, the mean rate for the 12 stimuli rated as being most natural was 194 syllables per minute, whereas the mean rate for the 12 stimuli rated as being least natural was 134 syllables per minute. Similarly, the mean naturalness rating of the 12 stimuli with the fastest speech rates was 4.9, whereas the mean naturalness rating for the 12 stimuli with the slowest speech rates was 6.8. This appears to indicate that listeners were relying more on rate than on the presence or absence of pauses to make judgments of speech naturalness.

GENERAL DISCUSSION

The purpose of these studies was to determine which factors of DECTalk synthesized speech contribute to perceptual ratings of highly natural and highly unnatural speech. The results from the first study are consistent with findings of other literature. Crabtree et al. (1990) and Mirenda and Beukelman (1990) documented that listeners preferred human speech to synthesized speech. Likewise, the listeners in study 1 rated human speech as being more natural than the synthesized speech. The subjective feedback from the listeners in the first study indicated that the factors of pitch, pause, and rate contributed most to the judgments of speech naturalness. However, the results from study 2 indicated that manipulation of pitch did not appear to be related to naturalness judgments. These results are similar to the findings of Venkatagiri (1991), who, unlike the studies reported here, used intelligibility as an independent variable. Venkatagiri (1991) found that pitch did not appear to consistently influence ratings of intelligibility. His findings, as well as those of study 2, appear to be at odds with the factors cited as important for making naturalness judgments by the listeners in study 1. It may be that, although average pitch did not seem to be a factor influencing naturalness ratings, variations in pitch may be a parameter that could influence ratings of naturalness. Pitch that has either too little (e.g., monotone speech) or too much variability may be rated more unnaturally. Future research that manipulates the acoustic characteristics of synthesized speech should explore this issue further.

The results of studies 2 and 3 revealed that manipulations of rate and pause did affect listeners' judgments of naturalness. Findings indicated that speech rates of 180 syllables per minute and above were rated as being more natural. These results are in contrast with the intelligibility/rate relationship results reported by Venkatagiri (1991), who found that decreasing the rate of syllables from 201 to 139 improved word intelligibility by 10%. One explanation of these differences may be attributable to the equipment used rather than the effects of speech rate judgments. Because the Echo II speech synthesizer used in the Venkatagiri study was from an earlier generation of synthesizers, differences in the results may reflect the influence of improvement in speech synthesis techniques. In addition, the current study controlled for intelligibility by providing listeners with the text of the stimuli, whereas Venkatagiri (1991) studied intelligibility itself and thus did not provide transcripts.

The findings of study 2 relative to rate are also not consistent with the findings of Higginbotham et al. (1994), that discourse text slowed to about 5.6 words per minute was summarized more completely. Again, task differences between the two studies may account for those divergent findings. In the Higginbotham et al. study, subjects listened to a paragraph to which they had not been previously exposed and were required to write down what they remembered. Having stimuli presented at such a slow rate may have facilitated the memory and processing components of this task. In the current study, subjects were given a hard copy of the text and were simply asked to rate naturalness. Because the listeners knew the text, and because they were not being asked to remember the content, they may have been less tolerant of slower rate.

Insertion of silent pauses in the synthesized speech output resulted in less naturally perceived speech. These data also were not consistent with those of Higginbotham et al. (1994), who found that listeners performed better on a task when the words were separated with 10-second pauses. However, the findings of

this study are consistent with Love and Jeffress (1971), who found that stutterers' fluent speech was perceptually different from the speech of normally fluent cohorts in that stutterers with fluent speech (following treatment) used a greater number of silent pauses compared to the speech of nonstutterers. In addition, the results of study 3 support previous reports of speech flow (i.e., fluency) as related to listeners' perception of naturalness in that stutterers' fluent speech (post-treatment) appears to be perceived by listeners as being more unnatural (Ingham et al., 1985; Metz, Schiavetti, & Sacco, 1990; Onslow et al., 1992).

Other factors may also have contributed to listeners' perceptions of reduced naturalness for stimuli with inserted pauses. Pauses that occur naturally may reflect the factors of formulation time for the speaker or a perceived need for processing time by the listener. However, the use of synthesized speech and the fact that transcripts were provided to ensure intelligibility may have created a situation in which the listeners were less tolerant of pauses and thus rated slower speech and speech with pauses as being less natural. In addition, the correlational analysis for study 3 suggests that it was not the pauses per se that affected perceived naturalness but the concomitant decrease in speech rate associated with more and longer pauses. Assessment of speech naturalness in situations closer to natural communication may help to clarify the role that pauses play in naturalness.

It is important to note that previous research was concerned with intelligibility, whereas the current set of studies addressed speech naturalness. Although a slow rate of speech presentation may facilitate processing time for a listener to make judgments of improved intelligibility, that same slow rate may also precipitate judgments of decreased naturalness. Thus, judging the naturalness of speech does not necessarily require the same resources as judging intelligibility.

Finally, the results of studies 2 and 3 suggest that although manipulating rate and pause consistently affected the ratings of naturalness, the magnitude of the observed differences was quite small (i.e., less than two scale points). The clinical implications of this finding are worth noting. Over the past decade, speech synthesis technology has become more sophisticated, cheaper, and more widely available (Hustad et al., 1998; O'Keefe, Brown, & Schuller, 1998; Reynolds, Bond, & Fucci, 1996). Clinically, it is well documented that voice output has a number of advantages over AAC systems without voice output. Someone using voice output can communicate from a distance without first having to obtain the attention of a partner. In addition, someone using a VOCA can communicate on the telephone and with individuals who are visually impaired and not literate (Higginbotham et al., 1994; McNaughton et al., 1994; Raghavendra & Allen, 1993). Although still an imperfect approximation of human speech, synthesized speech technology has reached a threshold whereby it is generally intelligible to a variety of listeners under a variety of conditions (Hustad et al., 1998; O'Keefe et al., 1998; Reynolds et al., 1996). This has allowed the users of VOCAs to become increasingly concerned with how their synthesized speech sounds. Indeed, VOCA manufacturers have provided a variety of parameters for customizing the voice output of such devices so that users can individualize synthetic speech to meet their needs and those of their listeners (Higginbotham et al., 1994; O'Keefe et al., 1998).

The results of this study indicate that the parameters of rate and pitch were identified by listeners as important in making naturalness judgments. However, manipulation of the rate parameter had only a modest impact at best on the perceived naturalness of synthetic speech, whereas manipulation of pitch had no impact. Additional research is needed to examine the influence of additional acoustic parameters and the use of more realistic communication situations on listeners' perceptions of naturalness.

Address reprint requests to: Ann Ratcliff, Central Michigan University, Moore Hall Room 441, Mt. Pleasant, MI 48859, USA; e-mail: Ann.e.ratcliff@cmich.edu.

REFERENCES

Baken, R. J. (1987). *Clinical measurement of speech and voice.* Boston: College Hill Press.

Coughlin, S. S. (1998). *Speech naturalness of normal speaking children and adolescents.* Unpublished doctoral dissertation, East Lansing, MI: Michigan State University.

Crabtree, M., Mirenda, P., & Beukelman, D. R. (1990). Age and gender preferences for synthetic and natural speech. *Augmentative and Alternative Communication, 5,* 256–261.

Finn, P., & Ingham, R. J. (1994). Stutterers' self-ratings of how natural speech sounds and feels. *Journal of Speech and Hearing Research, 37*, 326–340.

Gorenflo, C., Gorenflo, D., & Santer, S. (1994). Effects of synthetic speech, gender, and perceived similarity on attitudes toward the augmented communicator. *Augmentative and Alternative Communication, 13*, 87–91.

Guitar, B. (1998). *Stuttering: An integrated approach to its nature and treatment.* Baltimore: Williams and Wilkins.

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

Hustad, K., Kent, R., & Beukelman, D. (1998). DECTalk and MacinTalk speech synthesizers: Intelligibility differences for three listener groups. *Journal of Speech and Hearing Research, 41,* 744–752.

Ingham, R. J., Ingham, J. C., Onslow, M., & Finn, P. (1989). Stutterers' self-ratings of speech naturalness: Assessing effects and reliability. *Journal of Speech and Hearing Research, 32*, 419–431.

Ingham, R. J., Martin, R. R., Haroldson, S. K., Onslow, M., & Leney, M. (1985). Modification of listener-judged naturalness in the speech of stutterers. *Journal of Speech and Hearing Research, 28,* 495–504.

Ingham, R. J., & Onslow, M. (1985). Measurement and modification of speech naturalness during stuttering therapy. *Journal of Speech and Hearing Disorders, 50*, 261–281.

Ingham R. J., & Packman, A. C. (1978). Perceptual assessment of normalcy of speech following stuttering therapy. *Journal of Speech and Hearing Research, 21*, 63–73.

Kangas, K., & Lloyd, L. (1988). Early cognitive skills as to augmentative and alternative use: What are we waiting for? *Augmentative and Alternative Communication, 4*, 211–221.

Kreiman, J., Gerratt, B., Kempster, G., Erman, A., & Berke, G. (1993). Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36*, 21–40.

Love, L. R., & Jeffress, L. A. (1971). Identification of brief pauses in the fluent speech of stutterers and nonstutterers. *Journal of Speech and Hearing Research, 14*, 229–240.

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

Martin, R. R., Haroldson, S. K., & Triden, K. A. (1984). Stuttering and speech naturalness. *Journal of Speech and Hearing Disorders, 49*, 53–58.

Metz, D. C., Schiavetti, N., & Sacco, P. R. (1990). Acoustic and psychophysical dimensions of the perceived speech naturalness of nonstutterers and posttreatment stutterers. *Journal of Speech and Hearing Disorders, 55*, 516–525.

Mirenda, P., & Beukelman, D. (1990). A comparison of intelligibility among natural speech and seven speech synthesizers with listeners from three age groups. *Augmentative and Alternative Communication, 5*, 61–68.

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

Onslow, M., Hayes, B., Hutchins, L., & Newman, D. (1992). Speech naturalness and prolonged-speech treatments for stuttering: Further variables and data. *Journal of Speech and Hearing Research, 35*, 274–282.

Parrish, W. M. (1951). The concept of naturalness. *Quarterly Journal of Speech, 37*, 448–450.

Quist, R. W., & Blischak, D. M. (1992). Assistive communication devices: Call for specifications. *Augmentative and Alternative Communication, 8*, 312–317.

Raghavendra, P., & Allen, G. (1993). Comprehension of synthetic speech with three text-to-speech systems using a sentence verification paradigm. *Augmentative and Alternative Communication, 9,* 126–133.

Ratcliff, A., & Beukelman, D. (1995). Preprofessional preparation in augmentative and alternative communication: State of the art report. *Augmentative and Alternative Communication 11,* 61–73.

Reichle, J., York, J., & Sigafoos, J. (1994). *Implementing augmentative and alternative communication: Strategies for learners with severe disabilities.* Baltimore: Paul H. Brookes.

Reynolds, M., Bond, Z., & Fucci, D. (1996). Synthetic speech intelligibility: Comparison of native and non-native speakers of English. *Augmentative and Alternative Communication 12,* 32–36.

Reynolds, M., & Jefferson, L. (1999) Natural and synthetic speech comprehension comparison of children's two age groups. *Augmentative and Alternative Communication 15,* 174–182.

Runyan, C. M., & Adams, M. R. (1978). Perceptual study of the speech of "successfully therapeutized" stutterers. *Journal of Fluency Disorders, 3*, 25–39.

Runyan, C. M., & Adams, M. R. (1979). Unsophisticated judges' perceptual evaluations of the speech of "successfully treated" stutterers. *Journal of Fluency Disorders, 4,* 29–38.

Runyan, C. M., Bell, J. N., & Prosek, R. A. (1990). Speech naturalness ratings of treated stutterers. *Journal of Speech and Hearing Disorders, 55,* 434–438.

Sanders, W. C., Gramlich, C., & Levine, A. (1981). Naturalness of synthesized speech. In P. Suppes (Ed.), *University-level computer-assisted instructions at Stanford: 1968–80* (pp. 487–501). Stanford, CA: Stanford University.

Simpson, K., Beukelman, D., & Bird, A. (1998). Survey of school speech and language service provision to students with severe communication impairments in Nebraska. *Augmentative and Alternative Communication, 14,* 212–221.

Venkatagiri, H. S. (1991). Effects of rate and pitch variations on the intelligibility of synthesized speech. *Augmentative and Alternative Communication, 7,* 284–289.

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

APPENDIX

Directions to Participants

"We are studying what makes speech "natural" or "unnatural." You will be played 24 30-second audiotaped speech samples. Each sample will be introduced by the sample number, which you will write in the upper left corner of each response strip. Your task is to rate the speech naturalness of each sample. If the speech is highly natural to you, *circle* the number 1 ("highly natural") on that sample's response strip. If the speech is highly unnatural to you, *circle* the number 9 ("highly unnatural") on that sample's scale. If the speech is somewhere between highly natural and highly unnatural, *circle* the appropriate number on the scale. Do not hesitate to use the ends of the scale (1 or 9) when appropriate. Be sure to rate each sample. Naturalness will not be defined for you. Make your rating based on how natural or unnatural the speech is to you. You will hear each sample only once. Remember, however, that it is important that you number and rate each sample provided."