Talking Back to Big Bird: Preschool Users and a Simple Speech Recognition System

☐ Erik F. Strommen
Francine S. Frome

Strommen, E. F. & Frome, F. S. (1993). Talking back to Big Bird: Preschool users and a simple speech recognition system. Educational Technology Research and Development, 41, 5-16.

The present study examined the effectiveness of one configuration of automatic speech recognition (ASR) software and hardware with a child sample of 36 three-year-olds and a comparison sample of 20 adults. Subjects used a speaker-dependent, template-based system to play a simple "Sesame Street" naming game. Results indicated that while the system performed well with adults, it was much less effective with children. An analysis of the children's performances indicates that children's speech is more variable, in both volume and content, than that of adults. The ASR system responded ineffectively to this variability, resulting in inferior performance. Specific behaviors and their effects on the ASR system are identified, and possible system modifications that address these behaviors are noted.

☐ One of the most significant issues in the development of interactive educational products for young children is how best to design the software and hardware interfaces children must use to engage in the interaction itself. There is a growing body of evidence that young children have some difficulty with many of the hardware devices currently available (Char, 1990; Grover, 1986; Razavi, Medoff, & Strommen, 1991; Revelle & Strommen, 1990; Revelle, Strommen, & Offerman, 1990). It is therefore worthwhile to explore alternative methods of interaction that might be easier or more natural for children to utilize (cf. Hoko, 1986).

Advances in technology in the past few years have created one such alternative: speech (see Madlin, 1986, for examples of speech applications in adult employment contexts). Both as output and as input, speech has exciting implications as a feature of interactive learning materials for young children. It would be possible, for example, to design systems that teach pronunciation, that respond to spoken queries, or that emphasize abstract concepts, such as "alive," that cannot be easily communicated through icons and simple point-and-click methods. To capitalize on these possibilities, however, it is necessary to carefully examine how children engage and respond to speech-based interface systems, and to identify the features of such systems that

may need to be modified to make them effective for use by young children.

The existing literature on speech recognition is limited, and no studies of children's performance on such systems have been published. Nonetheless, existing studies give some indication of the potential difficulties young users might present to existing systems. In a comprehensive review of available studies, Simpson, McCauley, Roland, Ruth, and Willeges (1985) identified two problematic features of adult vocal behavior that have implications for the performance of young users. First, evidence from adult users indicates that speech under different conditions (e.g., stressful and nonstressful) varies in ways that reduce system accuracy. Rollins (1985) also found that variations in adult user pronunciation across quiet and noisy contexts impacted on recognition accuracy. Regardless of the conditions under which it is sampled, young children's speech is unlikely to be as uniform in cadence or pronunciation as that of adults. Speech systems that respond poorly to variations in adult speech may respond even more poorly to children's speech.

Second, Simpson et al. (1985) noted that extra words or sounds added to the speech sample that are unrelated to the required input also reduce system accuracy. It is not clear that young children are able to control their speech production in such a way as to carefully restrict the form of their responses only to utterances vital to the task.

An additional interface-related issue was observed by Rollins, Constantine, and Baker (1983) in their intensive assessment of four adult subjects using a voice recognition system at a job site. One subject commonly spoke when the system was not "listening," resulting in reduced system accuracy and effectiveness.

The above observations relate to inconsistencies of performance within individual adult users that have a negative impact on speech recognition systems and that are likely to be present in children's performance as well. There are also additional issues raised by the general effects of children's unique status as immature human beings acquiring a first language. A wide variety of studies have shown that children's articulation of speech differs in notable ways from that of adults (Gordon & Luper, 1989; Haelsig & Madison, 1986; Hubbard & Yairi, 1988; Mack & Lieberman, 1985; Nittrover, Studdert-Kennedy, & McGowan, 1989; Wijnen, 1988). How these differential performances impact on voice recognition systems is not known.

If designers of speech technologies are to create viable applications for young users, children's performances with such systems must be examined and problem areas identified. The purpose of the present study is to determine how children's speech performances, which differ from those of adults in notable ways, impact on the functionality of a simple speech recognition system. This article reports the results of an experiment designed to accomplish this goal by comparing the performance of a sample of three-year-old children to that of an adult control group performing the same simple speech-based task. By contrasting the performance of adults and children using the same vocabulary with the same speech system while performing the same task, the unique features of the performance of young children and the problems they present for speech recognition technology can be identified.

METHOD

Overview of the System

The platform for the technology in the study was a Toshiba 5200 MS-DOS computer with an add-on digital speech processing board. A clip-on microphone and portable amplifier (both battery-powered) were used as the voice input device. The speech recognition system utilized is based on a speaker-dependent model. Prior to actually using the system, the user "trains" the system to recognize the utterances needed in the interaction. The system "learns" to recognize a new word by encoding features of the sound into digital form and then analyzing them on a variety of dimensions. When sampling a new word, the system "listens" to any sound that reaches above a certain decibel threshold (in this case an amplified voice of 75 decibels against background noise) and keeps listening until the

sound energy drops below the threshold for approximately two-tenths of a second, at which point the system stops listening and regards the captured sample as an utterance. The utterance that was "heard" by the system is then analyzed, and a template based on the utterance's features is created. This template, once verified as reliable (see discussion below), is used by the system to recognize spoken responses.

The experimental task consisted of two separate steps, the *training sequence* and the *recognition sequence*. In the training sequence, the goal was to have subjects create a reliable template. The system assessed voice samples by comparing them on a variety of features and computing a summary measure of dissimilarity or "distance" between the features of previously learned templates and those of new samples as they were collected. A "reliable" template was created when two consecutive utterances identified as the same word fell within an arbitrarily specified distance range. If no two consecutive samples fell within the distance range, the last utterance for the target was used as the template.

After training was completed, the recognition sequence was executed immediately. In this task, five samples of the subjects' saying each of the four target words were collected one at a time, in a random sequence. Each sample was compared to the templates created for all four targets, and distances were computed for each comparison. The sample with the smallest distance between it and the template, and whose distance fell below the system's threshold, was taken as a valid match.

Participants

Two sample groups participated in the study: 36 preschoolers (M = 42.97 months or 3.6 years, SD = 3.11 months, *range* = 38–48 months) and 20 adults (M = 31.08 years, SD = 9.08 years, *range* = 21–56 years). Equal numbers of males and females were represented in both samples. The children were drawn from a single preschool in suburban New Jersey; the adults were volunteers drawn from Children's Television Workshop staff.

Procedure

The training and recognition sequences described above were embedded in a simple naming game in which children identified pictures of four "Sesame Street" Muppets (Big Bird, Cookie Monster, Elmo, and Grover) by saying their names. A previous study (Strommen, 1990) had found that these Muppets had high levels of consistent name recall among preschoolers, making them ideal candidates as stimuli for testing a system that both requires a large number of consistent verbal samples as inputs and must be motivating to young children.

Prior to actual use of the voice recognizer, a small clip-on microphone was clipped to the child's shirt or collar near his or her mouth. The child was told, "This is a talking game, and this is so the computer can hear you when you talk." The child was then shown pictures of the four Muppets in the study and asked to name each of them. Any child who did not know the Muppet's name was asked to make one up; if the child was reluctant to do so, the experimenter assigned the Muppet's name.

The child was then taken through a series of practice trials so as to become familiar with the task. At the start of the practice trial, the child was told, "Now, you watch the TV. When you see a color come on the TV, you say the name out loud." The child then moved through a series of colored screens in a press-button/see-color/say-name sequence until he or she completed the sequence accurately four times in a row. At this point the screen with the pictures of all four Muppets returned and the experimenter asked the names of the Muppets again, to ensure the child remembered them. The training sequence was then begun.

In training, the screen started out blank and then full-screen pictures of one of four Muppets (Big Bird, Cookie Monster, Elmo, and Grover) were shown to the child one at a time, in the same manner as the colors. The child was told, "Now, pictures of these people are going to show on the TV. When you see the picture, say the name of the person out loud." The Muppets appeared in a consistent order for each child, but the specific orders were var-

ied across children to avoid recency or primacy effects in the training sequence. The appearance of each picture was prompted by the experimenter's pressing a key, and the picture started the system "listening" for an utterance. After an utterance was "heard," the screen went blank to indicate completion of the listening phase. The Muppets cycled through the specified orders until, one by one, templates were created for them and they disappeared from the sequence. After a maximum possible 10 trials for each Muppet, the picture containing all four Muppets reappeared, and the experimenter then changed the screen to a 4 × 5 grid, with each Muppet appearing five times in random locations in the grid.

The appearance of the grid signalled the beginning of the recognition sequence. The child was told, "Now we're going to play a different game. In this game, we're going to make the people disappear, one by one, until they're all gone! Here's how we play: When I say go, you say the name of the person you want to disappear! Let's practice. Remember, when I say go, you say the name of the person you want to disappear. Ready? Go!" At this point, the child said the name of a Muppet, with the system inoperative. Any erroneous responses, such as saying several names in an unbroken string or saying "This one" and pointing rather than speaking, were corrected until performance reached a criterion of two appropriate responses in a row. The child then proceeded to follow this pattern with the system operative. After the child made each utterance, the system beeped to indicate completion of the sampling period. The experimenter then pressed a key to indicate which Muppet's name was spoken, or that an invalid utterance was spoken. The key indicating which Muppet's name was spoken caused one of the pictures of that Muppet to disappear. This routine was repeated until all the pictures were removed.

Additional prompts were used in the following situations:

1. In either training or recognition, if a child gave two inaudible responses in a row (i.e., the computer timed out or failed to register a response), the experimenter said, "The computer is having trouble hearing you. Try to talk a little louder." If the response following this prompt was still inaudible, the experimenter modeled talking louder for the child: "Say it nice and loud, like this: BIG BIRD! You try it."
2. In either the training or recognition sequence, if a child added extra words to the task by saying, "That's Big Bird" or "Big Bird now," the experimenter said, "Just say the person's name. Try not to say [the extra word]." This prompt was given twice, if needed.
3. In the recognition sequence, if a child was unable to understand the task or appeared reluctant to choose Muppets, the experimenter said, "I know! Let's play it this way: I'll point to a person and you tell me his name. Ready? Let's try it." The experimenter then pointed to each Muppet in turn and the child responded until the screen was cleared.

These same procedures were followed with the adult subjects, with slight modifications in the vocabulary of the instructions where appropriate to avoid "talking down" to them.

Dependent Measures

Inconsistent Spoken Responses

Speech recognition systems rely on consistent utterances from users; inconsistent utterances for the same target word cause errors in recognition. This variable was defined as the subject's either (a) using a new name for a Muppet (for example, calling Big Bird "yellow" after naming him correctly for several trials); (b) adding or subtracting syllables from a name; or (c) using completely different words (e.g., "Him again!").

Premature Onset of Phonation (POP)

This variable reflects the frequency of the subject's adding a nonword sound such as "Uh" or "Um" to the utterance prior to actually speaking the target word or phrase. Also referred to as a "filled pause," this behavior is thought to be related to the cognitive processing involved during speaking. There is no

consensus as to the exact nature of this behavior, but its implications for the disruption of automatic speech recognition are obvious (see Schachter, Christenfeld, Ravina, & Bilous, 1991, for a recent study of this behavior in adults).

System Cuts Off Sample after POP

This variable was scored if the system ended the sample collection period (i.e., stopped "listening") after the subject ceased the POP but before the intended utterance was actually spoken.

System Cuts Off Sample in Mid-Speech

This variable was scored if the system stopped "listening" while the subject was still saying the utterance.

Total Training Trials

Scored only for the training sequence, this variable is a measure of the difficulty of training the system to recognize the target utterances. A maximum of ten training trials per each of the four Muppets made a maximum score of 40 possible on this variable; since the system requires two consistent utterances to form a template, the minimum score was 8.

Average Distance Score

As described earlier, a numeric value reflecting the dissimilarity between two samples was computed, based on the speech recognition algorithm employed by the software. This number has no interpretable concrete value, but it serves as a useful indicator of the difference between two utterances as encoded by the system.

RESULTS

Scoring

The dependent variables in the present study were either collected by the computer during the task or scored from videotapes after the task was completed. All of these variables were standardized separately for the training and recognition sequences, using the following method. For each subject, the number of training trials on which a given behavior was observed was divided by the total number of training trials to yield a proportion of training trials on which the behavior was observed. The same procedure was followed for recognition trials. Since the number of training and recognition trials varied with each subject, this standardization was necessary to allow comparison of the performance across individuals, across the two samples, and across the training and recognition sequences. Summaries of the frequencies of all the variables for the training and recognition sequences are shown in Tables 1 and 2.

Each table presents two means: the full sample mean for each variable and the mean score exclusively for those subjects demonstrating a given behavior. The full sample (group) means indicate the level of the given measure for the age sample as a whole; the subsample means indicate the level of the given measure only for the subset of subjects in whom it occurred, so that the actual level of each performance can be assessed. For example, Table 1 indicates that the entire sample of children used inconsistent responses on only 2% of their trials ($M = 0.02$), but that the subset of children who actually exhibited this performance (22% of the total) made these errors on 9% of their trials ($M = 0.09$). Thus, this comparison shows that while the majority of children do not exhibit this problem, those few who do exhibit it do so at fairly high levels.

Subject-Based Variables

Broadly speaking, children's verbal responses to the system showed more variability than those of adults, as indicated by the results for the use of inconsistent names or words in both tables. This difference was tested using a 2 (sample) × 2 (sex) × 2 (mode: training or recognition) ANOVA on each variable. Results indicated that children tended to accidently use a different name or to say different words (e.g., "Him again!") more often than adults, $F(1, 52) = 8.90$, $p < .006$. The children also

TABLE 1 ☐ Summary of Training Trial Results by Sample

	THREE-YEAR-OLDS			ADULTS		
		Proportions of Trials			*Proportions of Trials*	
VARIABLE	% of *Sample*	Group Mean *(SD)*	Subsample Mean *(SD)*	% of *Sample*	Group Mean *(SD)*	Subsample Mean *(SD)*
Inconsistant names or words	22	0.02 (0.05)	0.09 (0.05)	–	–	–
Premature onset of phonation (POP)	61	0.13 (0.15)	0.20 (0.14)	15	0.02 (0.04)	0.12 (0.01)
System cuts off sample after POP	47	0.06 (0.08)	0.13 (0.06)	–	–	–
System cuts off sample mid-speech	58	0.11 (0.14)	0.18 (0.15)	35	0.05 (0.08)	0.13 (0.08)
Total trials	100	18.58 (7.41)		100	10.60 (2.48)	
Average distance*	64		992.54 (71.54)	95		753.00 (83.11)

Note: % of sample indicates number of children in sample demonstrating a given behavior. Group mean is calculated based on total sample size. Subsample mean is calculated based only on number of subjects manifesting behavior. Dash indicates behavior was not exhibitied by adult sample.

* Mean is calculated based on *n* subjects training all four Muppets.

TABLE 2 ☐ Summary of Recognition Trials Results by Sample

	THREE-YEAR-OLDS			ADULTS		
		Proportions of Trials			*Proportions of Trials*	
VARIABLE	% of *Sample*	Group Mean *(SD)*	Subsample Mean *(SD)*	% of *Sample*	Group Mean *(SD)*	Subsample Mean *(SD)*
Inconsistant names or words	31	0.03 (0.05)	0.07 (0.05)	–	–	–
Premature onset of phonation (POP)	73	0.19 (0.21)	0.26 (0.20)	15	0.01 (0.02)	0.04 (0.002)
System cuts off sample after POP	73	0.10 (0.10)	0.13 (0.10)	5	0.002 (0.01)	0.04
System cuts off sample mid-speech	65	0.08 (0.10)	0.11 (0.10)	15	0.01 (0.03)	0.08 (0.04)
% correct recogni-tions*	64		48.23 (22.31)	95		88.26 (9.14)
Average distance**	41		1033.25 (73.62)	95		797.01 (72.14)

Note: % of sample indicates number of children in sample demonstrating a given behavior. Group mean is calculated based on total sample size. Subsample mean is calculated based only on number of subjects manifesting behavior. Dash indicates behavior was not exhibited by adult sample.

*% correct recognitions based on subjects with four valid templates only.

**Average distance is calculated based on *n* of subjects training all four Muppets.

exhibited significantly higher levels of premature onset of phonation (POP)—the tendency to say "Um" or "Uh" prior to speaking—than adults, $F(1, 52) = 23.58$, $p < .0001$. There were no gender or mode effects and no significant interactions for either variable.

The significant differences between the samples are important not only because the inclusion of extra words affects the creation and recognition of templates in the system, but also because the children appeared to have much less ability to monitor and correct their performance than adults. While adults typically required one prompt to stop POP, many children continued their POP even after two prompts. While children were better at controlling the occasional use of extra words, such as "That's Grover!" or "Grover again!" the need to monitor their speech was clearly an extra demand on their performance with which they found it difficult to comply.

System-Based Variables

The tendency of children to demonstrate POP as part of their performance affects the system in a way that uniquely impacts on its performance when responding to them. The system cut off the sample significantly more often when children paused between the premature phonation and the actual utterance than it did with adults, $F(1, 52) = 23.75$, $p < .0001$. There were no other significant effects. The practical result was that when the child said something like "Um. . . Elmo!" the system accepted "Um" as the sample rather than "Elmo!" The consistency of this kind of performance was such that, for several children, it appeared that for certain templates the system was actually trained to the child's premature phonation rather than the name of the Muppet.

Even when subjects did not demonstrate POP, in both modes the computer cut off the sample in mid-speech significantly more often for children than for adults. This was more likely to occur for the names Big Bird and Cookie Monster. The ANOVA yielded a significant effect for sample, $F(1, 52) = 6.05$, $p < .02$, $M = 0.10$ of trials for children and $M = 0.03$ of trials for adults. There was also a significant effect for mode, $F(1, 52) = 5.82$, $p < .02$, $M = 0.09$ of trials during training, $M = 0.05$ during recognition. No other effects were significant.

Training Performance

The subjects in the present study were given a maximum of ten trials for each Muppet in which to train a template. If a reliable template was not generated after the tenth trial, the subject was scored as having failed to train a reliable template. There was a striking difference between children and adults in the ability to train the system. Only one adult failed to train the system, and that was only for a single Muppet. In contrast, 13 children, or 36% of the sample, failed to train the system on at least one Muppet, a significant difference between the samples $\chi^2(1, n = 56) = 5.08$, $p < .02$. Out of these 13 children, 9 failed to train one Muppet, 3 failed to train two, and one child failed to train three of the four Muppets (Big Bird was the only one trained). Seven children failed to train the Cookie Monster template, five failed to train the Grover template, and three failed to train the Big Bird and Elmo templates.

The subject- and system-based variables noted above are correlated with varying degrees of strength with the total number of trials children required to train the system: using inconsistent names or words, $r(36) = .30$, $p < .08$; POP, $r(36) = .59$, $p < .0001$; cutting off the sample after a POP, $r(36) = .44$, $p < .008$; and cutting off a sample in mid-speech, $r(36) = .29$, $p < .09$. While the relationships vary in strength, all are positive—the more a given behavior is performed, the more trials are required to form a template, if one is formed at all. This finding suggests that these variables impact on template creation, most likely through the mechanism of producing unacceptable samples to be matched to one another during the training sequence.

Distance Scores

As mentioned earlier, the distance measure serves as the criterion of dissimilarity between templates and samples. Unfortunately, data regarding distances must be restricted to those subjects producing reliable templates, because a distance calculation that fell outside the set threshold always defaulted to an arbitrary maximum value, meaning that failed training trials produced no useful distance results. However, a 2 (sample) × 2 (sex) × 4 (Muppet) ANOVA of the distances computed for template matches for adults and children who created four valid templates did indicate a significant sample main effect, $F(1, 38) = 109.79$, $p < .0001$, as well as a significant main effect for Muppet, $F(3, 114) = 5.09$, $p < .002$.

The pattern of differences indicates two general findings: First, adult distances were consistently lower than those of children by several hundred units for each target, despite the fact that all subjects included in the analysis trained the system effectively. Second, the distance results indicate significant differences among the target phrases, such that, for both age levels, Cookie Monster was the most difficult template and Elmo was the easiest. This effect for Cookie Monster and Elmo across both age groups suggests that the vocabulary chosen for the task is important to how well the ASR system functions.

There was also a small but significant sex × sample interaction for distance, $F(1, 38) = 5.31$, $p < .027$, such that three-year-old girls obtained lower distance scores than did three-year-old boys ($M = 959.64$ for girls and $M = 1022.71$ for boys), but adult women obtained slightly higher distance scores than did adult men ($M = 772.98$ for women and $M = 730.81$ for men). These differences, while statistically significant, are so small in terms of the actual scale of the distance units that they have no practical value and will not be considered further.

Finally, none of the performance variables noted above showed significant relationships with the distances computed in successful training trials, suggesting that distance scores may have been influenced by some other specific features of the collected voice samples related to articulation quality.

Differences between Successful and Unsuccessful Children

A comparison of the children who succeeded in creating reliable templates for all four Muppets ($n = 23$) with those who trained three or less ($n = 13$) underscores the above results. Children who failed to train all four templates took significantly more trials in training ($M = 25.77$) than those who did train all of the templates ($M = 14.52$), $t(34) = 6.41$, $p < .0001$. (Note: Optimal performance is 8 trials.) The only statistically significant difference between the groups on the variables described above is that children who failed to create one or more templates exhibited a significantly higher number of POPs ($M = 0.20$) than those who trained all four ($M = 0.09$), $t(34) = 2.21$, $p < .034$. However, while the differences did not reach statistical significance, the children who failed to train all four templates also demonstrated higher levels of the system's cutting off the sample after the POP and cutting off the sample in mid-speech.

Differences between Successful Children and Adults

A comparison of the three-year-olds who trained four reliable templates ($n = 23$) with adults who did so ($n = 19$) indicates that the children are still more like their agemates than the adults. Results from a 2 (sample) × 2 (sex) × 4 (Muppet) ANOVA of number of trials to creation of a reliable template indicate that the only significant effect is for sample, $F(1, 38) = 12.47$, $p < .001$, and is in the expected direction: The successful children required approximately twice as many trials as adults to create a valid template. The higher number of trials can be related to the children's exhibiting higher levels of the behaviors noted above than their adult counterparts. The children demonstrated POP significantly more often, $t(40) = 3.06$, $p < .004$ ($M = 0.09$ for children vs. $M =$

0.02 for adults) and, although the differences are not statistically significant, the computer cut off the sample in mid-speech more often (M = 0.10 for children vs. 0.05 for adults) and after their POP as well (M = 0.06 for children vs. M = 0 for adults). Lastly, while none of the adults used different names or referential utterances such as "That one," five out of the 23 children did so at least once.

Effects of Performance on Recognition

The purpose of the recognition sequence was to test the system's accuracy in matching Muppet names produced in the "game" context to the templates created by the ASR during training. Each Muppet name was collected five times and each sample was matched against the four templates that were created in the training sequence. As previously described, if a reliable template was not created, the ASR attempted to match the utterance against the last sample that the system had stored for that Muppet. In these cases, it is theoretically possible for the system to make such a match; in the present data, however, out of 95 possible matches against invalid templates, only one occurred (for Cookie Monster). For this reason, the analyses of the system's accuracy will be restricted to those 42 subjects (23 three-year-olds and 19 adults) who trained four reliable templates.

A 2 (sample) × 2 (sex) × 4 (Muppet) ANOVA on total correct matches revealed a significant effect for sample, $F(1,38) = 50.36$, $p < .0001$; for Muppet, $F(3, 114) = 6.91$, $p < .0001$; and for the sample × Muppet interaction, $F(3, 114) = 5.05$, $p < .003$. The interaction effect appears to be due to the fact that the difference between the three-year-olds' and adults' mean success rates for Big Bird is much smaller than for the other three Muppets (a difference of 0.87 for Big Bird, but 2.19 for Cookie Monster, 2.22 for Elmo, and 2.70 for Grover).

The difference between samples in successful matches is as striking as that obtained for trials in the training sequence. With the exception of Big Bird, the ASR's success rate with three-year-olds is approximately half of its success with adults: for Big Bird, M = 3.34 for children vs. M = 4.21 for adults; but M = 1.65 vs. M = 3.84 for Cookie Monster, M = 2.57 vs. M = 4.79 for Elmo, and M = 2.09 vs. M = 4.79 for Grover. The final translation of these numbers into an average successful recognition rate varies according to how the system is evaluated. If all children and adults are considered, regardless of the number of valid templates they created, then out of each person's total possible matches, the ASR was successful with children 42% of the time (M = 8.4 out of 20 trials), and with adults 89% of the time (M = 17.8 out of 20 trials). If only subjects who created four valid templates are considered, the success rate for the three-year-olds is 48% and the success rate for the adults is unchanged.

The system's substantially lower success rate with three-year-olds is not due to false matches between the child's utterance and the template for a different Muppet name: only six out of 317 matches were erroneous, a false match rate of less than 2%. The system's low success rate is clearly due to its inability to match child voice samples with their templates.

The same subject- and system-based variables reported earlier for the training sequence have an impact on recognition, but their effects are not reflected in the recognition accuracy rate. In the recognition sequence, the experimenter was able to discard trials where a different name or phrase was spoken, or trials where a POP was accepted as the sample; these cases were not counted as trials in the success rate analyses above.

The relationships between recognition accuracy and dependent measures that do reflect on the collected sample are similar to those noted above for training, and also indicate that poor performance in the training sequence impacts negatively on later recognition accuracy. For example, cutting off the sample in mid-speech during both training and recognition is negatively related to the accuracy score for three-year-olds, $r(36) = -.35$, $p < .037$ for recognition; $r(36) = -.49$, $p < .001$ for training. POPs during recognition were treated as invalid trials and were not counted as part of the five accuracy trials for each Muppet; how-

ever, POPs during training are strongly related to recognition accuracy, $r(36) = -.52$, $p < .0001$. The relationship between POPs during training and recognition accuracy suggests that the templates created with POPs as part of the samples are of poor quality.

Other Behaviors

Although only noted in one child in the present sample, children speaking with their fingers in their mouths should be considered a potential problem, since this behavior will definitely reduce the quality of the utterance. The single subject in this study who was resistant to requests to stop putting his fingers in his mouth was also the only child who failed to train three out of the four templates.

Another behavior in children that was almost unique to recognition and that also deserves attention is pointing rather than speaking during the game. The frequency of pointing showed a significant effect for sample, $F(1,53) = 8.67$, $p < .005$; and a significant sample × mode interaction, $F(1, 53) = 7.95$, $p < .007$. These results are due to the fact that while no adults pointed to the Muppets as they answered (or instead of answering, which was more common), 11% of the children did so during the training sequence and 36% did so during the recognition sequence. This tendency toward nonverbal responding on the part of the children caused time-outs and inaudible trials and is related to the use of different utterances, such as saying "That one!" instead of the appropriate name.

DISCUSSION

The results of the present study indicate that a voice recognition system that works well with adults can nonetheless be ineffective and problematic with young children. Young children appear to present two distinct sets of problems to current speech technologies. First, children's vocal performance has a variety of dimensions that impede system functioning. Second, it appears that children's immature articulation presents voice processing systems with an unexpectedly severe challenge. It would appear that both technical and interface-related innovations will be required if the promise of voice interfaces for child users is to be realized.

Children in the present study demonstrated significant variability in their vocal performance. They included extra words and premature phonations at levels significantly higher than those of adults. Even children who trained the system effectively showed levels of these behaviors higher than those of the adult sample. The results of these behaviors was the production of poor quality templates for recognition and a dramatic decrease in system accuracy. In addition, children's speech appears to vary more in volume level and contain longer pauses than that of adults, as suggested by the significantly higher frequency of trials where the system cut off speech samples for the children than for the adults. The premature termination of sampling in the midst of children's utterances had predictable effects on both the number of samples required to create reliable templates and on recognition accuracy.

It is possible that the issues of sample loudness and utterance variability can be addressed through modifications in the technology itself. Changes in decibel thresholds and in the length of silent intervals the system tolerates seem likely as solutions. However, to the extent that children's performances will continue to require extra trials or require the children themselves to modify their performance, new interface structures will be required. Studies of how automated systems can best obtain modified performances from children are clearly needed. Similarly, basic research on how best to instruct children to modify their vocal performance, using technology or not, is also needed. The current speech literature focuses almost exclusively on the therapeutic remediation of language-impaired children, not on how linguistically average children can be taught to change their speech style or patterns.

The second major finding in the present study is that children's immature articulation

seems to present unexpected difficulties to voice processing systems. The distance computations produced for children's utterances, which are an index of the ability of the ASR to process them, were significantly higher than those of adults. This was the case even for children who were successful at both training and recognition, suggesting a general problem with children's speech production.

It was not possible to determine which components of children's articulation were the most difficult for the system. Future studies must give this issue close scrutiny. One important finding is the main effect for Muppet name on distance computations. This effect suggests that the vocabulary used in the system has a significant influence on its effectiveness. While the nonrepresentative set of words deployed in the present study prevents us from drawing any conclusions about the type of word sounds best suited for this task, more systematic research will likely be able to identify particular words children can employ that are maximally identifiable by speech systems and that produce minimal problems for processing.

An additional way to solve this problem could be to determine whether alternative systems for speech recognition currently under development are more accurate than the one used in the present study when children are the users. The present study utilized a speaker-dependent, template-based methodology. Speaker-independent systems, or systems that rely on Markov models to compare templates, may be superior for young children, regardless of vocabulary content. Word-spotting systems which can detect the target phrase in a string of unrelated speech could also be especially helpful.

In conclusion, speech interfaces hold great potential for empowering young children in their use of interactive technology. However, it is apparent that children bring a variety of unique behaviors and developmental limitations to the human-computer dialogue. These limitations severely reduced the effectiveness of the speech system in the present study, and there is reason to believe they will have a similar impact on other systems. In order to capitalize most effectively on this technology's power for education, the unique requirements of child users must be carefully identified. Specific technological and interface modifications that can facilitate children's use of speech systems to their fullest potential will have to be designed and tested. Such an undertaking will not only result in more effective speech systems, but will provide important new data to the fields of developmental psychology, human factors, cognitive science, and language research. Perhaps most important, however, is that ultimately such efforts will expand the community of users of technology to include a group too often overlooked: its youngest members. □

Erik F. Strommen is with Children's Television Workshop and Francine S. Frome is with AT&T Bell Laboratories.

REFERENCES

Char, C. (1990, April). *Touch, click, or jump? The relative merits of different input devices for use by young children*. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Gordon, P. A., & Luper, H. L. (1989). Speech disfluencies in nonstutterers: Syntactic complexity and production task effects. *Journal of Fluency Disorders, 14*, 429–443.

Grover, S. C. (1986). A field study of the use of cognitive-developmental principles in microcomputer design for young children. *Journal of Educational Research, 79*, 325–332.

Haelsig, P. C., & Madison, C. L. (1986). A study of phonological processes exhibited by 3-, 4-, and 5-year-old children. *Language, Speech, and Hearing Services in the Schools, 17*(2), 107–114.

Hoko, J. (1986, Nov-Dec). Alternatives to keyboarding. *Tech Trends*, 23–24.

Hubbard, C. P., & Yairi, E. (1988). Clustering of disfluencies in the speech of stuttering and nonstuttering preschool children. *Journal of Speech and Hearing Research, 31*, 228–233.

Mack, M., & Lieberman, P. (1985). Acoustic analysis of words produced by a child from 46 to 149 weeks. *Journal of Child Language, 12*, 527–550.

Madlin, N. (1986, April). Conversant computers. *Management Review, 75*, 59–60.

Nittrover, S., Studdert-Kennedy, M., & McGowan, R. S. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-

vowel syllables spoken by children and adults. *Journal of Speech and Hearing Research, 32*, 120–132.

Razavi, S., Medoff, L., & Strommen, E. (1991, April). *"Do I push this button?": Three year olds' use of the Nintendo controller.* Paper presented at the annual meeting of the Eastern Psychological Association, New York.

Revelle, G. L., & Strommen, E. F. (1990). The effects of practice and input device used on young children's computer control. *Journal of Computing in Childhood Education, 2*(1), 33–41.

Revelle, G., Strommen, E., & Offerman, S. (1990). *Cursor and device differences in children's computer control.* Manuscript submitted for publication.

Rollins, A. (1985). Speech recognition and manner of speaking in noise and in quiet. In L. Borman & B. Curtis (Eds.), *CHI '85 Conference Proceedings* (pp. 197–199). Reading, MA: Addison-Wesley.

Rollins, A., Constantine, B., Baker, S. (1983). Speech recognition at two field sites. In A. Janda (Ed.), *CHI '83 Conference Proceedings* (pp. 267–273). Reading, MA: Addison-Wesley.

Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology, 60*, 362–267.

Simpson, C. A., McCauley, M. E., Roland, E. F., Rutch, J. C., & Willeges, B. H. (1985). System design for speech recognition and generation. *Human Factors, 27*(2), 115–141.

Strommen, E. (1990). *Consistency and variability in children's naming of Muppet characters.* Unpublished manuscript, Children's Television Workshop.

Wijnen, F. (1988). Spontaneous word fragmentation in children: Evidence for the syllable as a unit of speech production. *Journal of Phonetics, 16*, 187–202.