Designing Visualizations to Facilitate Multisyllabic Speech with Children with Autism and Speech Delays

Joshua Hailpern, Andrew Harris, Reed La Botz, Brianna Birman, Karrie Karahalios
Department of Computer Science
University of Illinois at Urbana Champaign
jhailpe2@cs.uiuc.edu, {harris78,labotz1,birman1}@illinois.edu, kkarahal@cs.uiuc.edu

ABSTRACT

The ability of children to combine syllables represents an important developmental milestone. This ability is often delayed or impaired in a variety of clinical groups, including children with autism spectrum disorders (ASD) and speech delays (SPD). Prior work has demonstrated successful use of computer-based voice visualizations to facilitate speech production and vocalization in children with and without ASD/SPD. While prior work has focused on increasing frequency of speech-like vocalizations or accuracy of speech sound production, we believe that there is a potential new direction of research: exploration of real-time visualizations to shape multisyllabic speech. Over two years we developed VocSyl, a real-time voice visualization system. Rather than building visualizations based on what adult clinicians and software designers may think is needed, we designed VocSyl using the Task Centered User Interface Design (TCUID) methodology throughout the design process. Children with ASD and SPD, targeted users of the software, were directly involved in the development process, allowing us to focus on what these children demonstrate they require. This paper presents the results of our TCUID design cycle of VocSyl, as well as design guidelines for future work with children with ASD and SPD.

Author Keywords

Visualization, Autism, Children, Syllable, Speech

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: Screen design, Voice I/O; K.4.2 Social Issues: Assistive technologies for persons with disabilities

General Terms

Experimentation, Human Factors.

INTRODUCTION

The process of developing language is "a unique characteristic of human behavior. . . [that] contributes in a major way to human thought and reasoning" [26]. Moreover, the ability to combine syllables is a critical milestone in speech development [17, 42], and is key to optimizing language growth. However, some children with **A**utism **S**pectrum **D**isorders (**ASD**)[1] and **Sp**eech **D**elays (**SPD**) do not develop language on their own from the input typically presented to them (i.e., via the auditory modality [27]). Hence, the natural development of social behavior, such as spoken communication, is significantly disrupted. This results in detrimental effects on many aspects of their lives [17, 42]. We therefore sought alternative forms of presentation and feedback to facilitate their learning to communicate. Specifically, we aim to develop solutions to help teach multisyllabic word production: either as word combinations (e.g., "more juice") or as individual multisyllabic words (e.g., "banana").

Visual information has emerged as a critical form of support for children with ASD due, in part, to documented strengths of this form of processing information [33]. Further, the use of technology has the potential to alleviate some apprehension experienced by many ASD and SPD children when interacting with people [3]. Given the success of researchers in using visualizations to encourage vocalization [14], we believe that voice visualizations on the computer will provide therapists with new techniques to enhance existing approaches.

Visualizing voice represents a vast design space. Which design considerations are essential? Due to language challenges, we cannot do interviews with users with ASD/SPD to ask "how do you like existing systems? Therefore, to develop these software tools, we formed a team from Computer Science, Speech and Hearing Science and Special Education, and employed a **T**ask **C**entered **U**ser **I**nterface **D**esign (**TCUID**) methodology [25], in which subjects drawn from the target populations were involved with software development throughout the design phase. In effect, children with ASD and SPD became part of the design team. Because children with ASD and SPD have difficulties communicating, we examined their preferences as well as their interactions with the researcher and technology. To further inform our design, we included children without speech-language delays to provide explicit verbal feedback about the software, which can further uncover usability challenges.

Our main contributions are the the documentation of the TCUID process, the resulting design of VocSyl (software and visual-

DIS 2012, June 11–15, 2012, Newcastle, UK.

[1] The prevalence of ASD is estimated by the Center of Disease Control and Prevention (CDC) to be 1 in 150 children [6]

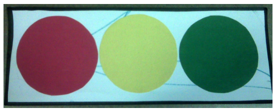


Figure 1. Pacing Board for 3 Syllable Word

Paper, with three construction paper circles glued down.

izations), and the articulation of design guidelines for future software development that shapes multisyllabic production (that are deeply grounded in observations of children with ASD/SPD). We begin with a review of the literature at the intersection of HCI, speech therapy, and ASD/SPD. Using prior work to guide our system, we next present the system and architectural features of VocSyl and describe how we ensured flexible and rich real-time visualizations. We then discuss the *TCUID Study*, including both the methodology used and the design changes made to our software based on the interactions in the design process. Our findings from the past two years are summarized in a set of *Design Guidelines* for future research and HCI software design. Finally, we conclude with a discussion of an ongoing intervention study to examine the impact of our software.

BACKGROUND & LITERATURE REVIEW

For all children, multisyllabic speech represents a critical milestone – aiding the communication of more sophisticated concepts through phrase development and phonologically complex words [44]. Challenges with syllable combinations (within or across words) can impact a children from late-talking toddlers to children with apraxia of speech [17, 42]. In addition, there is indication that one out of every three children with ASD do not develop functional speech at all[2] [5]. Difficulties in multisyllabic productions can persist and result in other phonological difficulties later in life [36]. Given the importance of multisyllabic speech, it is striking how scant the literature remains in regard to related interventions. This section will focus on speech and technology interventions. We focus on the ASD/SPD populations due to evidence of their difficulties in developing intelligible spoken language and evidence of the effectiveness of visual aids.

Therapy for Children with Autism & Speech Delays

Many of the characteristic difficulties experienced by children with ASD revolve around communication, empathy, social functioning, and expression. Since the 1960s, Ivar Lovaas' pioneering and successful approach of **A**pplied **B**ehavior **A**nalysis (**ABA**) has been one of the main methods to teach communication and social skills [26]. However, frequent therapy sessions that require sustained attention and intense human-to-human contact can produce anxiety, which can cause difficulty for both practitioners and children with ASD [20]. This motivated us to explore technology-based solutions which have the potential to minimize anxiety [3], improve language skills, and may increase the willingness of children to participate in the therapy.

In addition, providing information through the visual modality has repeatedly been documented as a useful treatment approach for children with ASD [33]. For example, tools like the Picture Exchange System [4] and visual schedules [8] have been widely adopted to support interaction and day-to-day functioning. However, the use of visual feedback to provide real-time acoustic information about vocal productions for children with ASD is rare[3]. In addition, few if any interventions for children with ASD have focused explicitly on multisyllabic productions. In contrast, a few treatment studies targeting multisyllabic productions have been published in relation to children with specific speech-language delays (e.g., [40]). The most relevant approach is the use of a static visual display known as a pacing board [44]. The pacing board, represented in Figure 1, provides a graphic representation of individual syllables via circles. As a word is practiced, the clinician and child touch each circle as its corresponding syllable is produced. However, the pacing board is neither dynamic, nor does it provide acoustic information about the child's or clinician's vocal productions.

Technology Research on Autism and Speech Delays

Computers have been shown to be an important platform for interventions for children with ASD. Researchers have examined the role of technology in early diagnosis [23], teaching human-to-human interaction to high-functioning children with ASD [22, 43] and reducing the apprehension caused by human-to-human interaction in play [31]. Morris et al., have examined the customizability of computer systems to provide a more inviting learning environment for children with ASD [34]. However, the majority of this research does not focus on speech acquisition and speech skills.

There is a vast design space of voice-based technology and research systems. Of note, HCI research [11, 19] and behavioral research [2, 38] have used computer solutions in the context of speech therapy and communication. Within the ASD community, Hoque has explored sentence practice with higher functioning individuals with ASD through the use of games [18]. Nonetheless, no prior work could be found that examined the role of technology in teaching multisyllabic speech. One potential reason is that many designers using technology focus on lower and higher skill levels: sentences for high-functioning children [18] or vocalizations for nonverbal functioning children [14]. Because of limitations in speech recognition software [35, 41], forms of speech detection are limited, especially for individuals with poor diction.

SCOPE & MOTIVATION

Our research explores the process of building computer systems to aid in shaping multisyllabic speech in children with ASD and SPD, specifically by including these children, who have limited speech-language skills, in the design process. Because this design space is so large, and teaching multisyllabic speech to children with ASD/SPD using technology is a novel interaction, launching straight into a validation study of our software would be premature (see Future Work for ongoing studies). Further, prior work has repeatedly shown that lessons learned from TCUID, loosely structured studies and case studies[4] are an important contribution

[2] Speech development is consistent with other forms of communication, such as signs or picture symbols [32].

[3] SpeechViewer[19] and VisiPitch[21] are noteworthy exceptions, though they do not target teaching multi-syllabic productions.

[4] Even without explicit validation studies.

[10, 7]. They improve development time, usability, and help designers better understand context and users [45]. In this spirit, we wish to share our findings to help others working in a similar domain. Given the novelty of this design space, other designers can greatly benefit from the numerous technical and user-centered lessons learned, which we were only able to uncover through a TCUID investigation.

Our main contributions are the documentation of the TCUID process, the resulting design of VocSyl, and the design guidelines for future software development that shapes complex speech. Our research bridges the gap between Hailpern's work [14] on encouraging vocalization in children and Hoque's work [18] on providing a game to practice full sentence production. Note that teaching perfect speech or precise phonemes is outside the scope of this research: our software focuses on facilitating prosody and syllables. We employed well accepted **D**igital **S**ignal **P**rocessing **(DSP)** algorithms so as to creatively engage our participants and assess interaction [5].

VOCSYL SOFTWARE

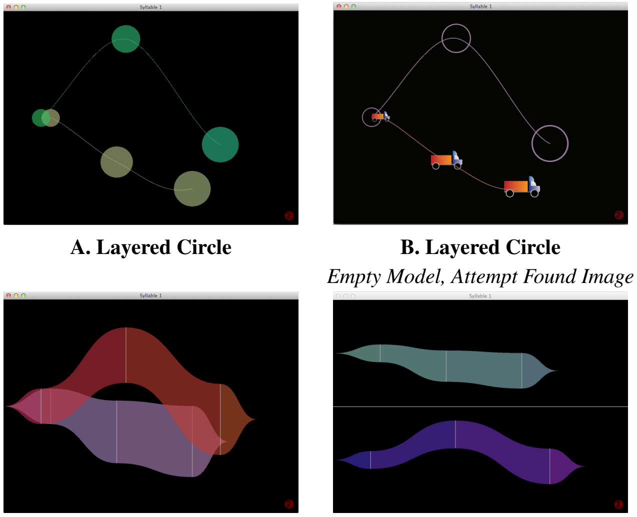
VocSyl[6] is a Java based, real-time audio visualization system for use within a clinical speech-therapy setting. It utilizes the Processing API to render graphics. To teach multisyllabic speech production, we visualize changes in syllables, timing (speed), tone (pitch) and emphasis (volume). Figure 2 illustrates four of the many VocSyl visualizations.

Research software is often designed and built with little consideration for the software's use beyond the conclusion of the experiment. This approach is effective for studying how humans interact with software, but does not yield software that is usable outside of a research setting. In contrast, we built VocSyl to be an extensible software architecture that could be used on a regular basis in a clinical setting.

Use Case

VocSyl is based on a common form of ABA behavioral training [1]. The clinician first says a word. This prompt is called a **model**. The clinician then waits for a response from the subject. We term this response an **attempt**. If the attempt approximates the model, the subject is given a reward (e.g., food or verbal praise). If the attempt does not match the model, the trial is repeated. When the clinician says a target word or phrase, the auditory information is fleeting and does not persist. Therefore, if a child says a word incorrectly, there is no lasting product presented to the child for comparison (aside from playing back a recording of their voice).

VocSyl aims to present a visual representation (see Figure 2) of the vocal features of a word attempt. When the clinician prompts the child with a word, a visual representation of the uttered word appears in near real time on a screen. When it is the child's turn to attempt the word, their utterance is "drawn" in real time next to, or on top of, the visualization of the model. As a result of VocSyl's persistent visualization of auditory features, the child and clinician can visually compare the differences between the model and the attempt.


A. Layered Circle

B. Layered Circle
Empty Model, Attempt Found Image

C. Layered Envelopes

D. Split Screen Envelopes

Figure 2. Four Examples of VocSyl Visualizations
These examples are of two utterances of the word "basketball," the first where each syllable is said in descending pitch, and the second where the second syllable is said at a higher pitch than the first and third syllables.

Real-Time Multisyllabic Speech Visualization

This visualization has two basic **styles** loosely based on the pacing board commonly used in speech therapy (Figure 1). We term the style *circle* to refer to the shapes in Figure 2 A and B (where each syllable is an individual circle). We term the style *envelope* to refer to the shape in Figure 2 C and D (where each syllable is a different segment separated by a vertical white line). The circle visualization can be modified. The thin line smoothly connecting the circles together (Figure 2 A and B) can also be turned on and off. In addition, the visualization of the model has the option to be an empty circle (simulating a target to hit and match as in Figure 2 B).

This VocSyl visualization dynamically captures and shows four key elements of each utterance. *Syllables* are represented by discrete elements on the screen (envelope or circle). *Pitch or tonal changes* (relative to each user's starting pitch) are illustrated on the y-axis[7]. *Emphasis* or stress is represented by envelope or circle size (thickness or diameter respectively). *Pacing/timing* is represented on the x-axis.

Real-Time Audio Analysis

Systems that rely upon real-time visualizations, like VocSyl, require both visual rendering and audio processing. Because synchronicity is central to real-time visualizations, if one element of this system (audio analysis or rendering) dominates the applications' execution, the whole system suffers and will not run in real time. To this end we followed the *Publish/Subscribe* design pattern [13]. This provides a robust back end that will not deadlock. It scales well as we add

[5] Thus, designing new, improving existing or using state-of-the-art DSP algorithms is outside of the scope and goal of this type of work (similar to the idea of interpretative affordances [24, 29]).

[6] VocSyl's name is an amalgam of the words **Voc**alization and **Syl**lable, acknowledging its use, while its pronunciation (Voxel / *vaksl* /) is a nod to its use of visualizations and computer graphics.

[7] We used relative change because the voice of an adult is lower than that of a child. Relative change in cents is a linear delta between starting pitch and current pitch (unlike Hz which is logarithmic).

more concurrently running complex analysis, while maintaining smooth real-time visualizations.

The VocSyl system can analyze live microphone audio, and pre-recorded WAV files (for user practice, without a clinician present). To perform real-time DSP, we have implemented a series of band-pass filters (BPF)[8]. The main DSP of VocSyl revolves around prosody: syllables (building blocks of words), pitch (intonation), time (tempo), and volume (emphasis). We therefore built components that can extract audio's volume (in amplitude or decibels), pitch (calculated using a sliding window Cepstrum Plot [37]), and syllable occurrence (using convex-hull analysis [30]). Adding new DSP is a simple process (covered in the next subsection).

Plugin Architecture for Expandability

VocSyl's modular architecture allows the system to be highly extensible. Developers can easily add new DSP or visualization by building on top of the existing components. A new visualization can be created in roughly 100 lines of code using pre-built components that detect volume, pitch and syllables. While the specific visualizations used in this research are more complex (and required more code), our framework allows future visualization developers to focus on *how* to represent the vocal properties, and not the complexity of the underlying VocSyl system.

TASK CENTERED USER INTERFACE DESIGN

Once the VocSyl architecture was completed, we began TCUID. Over a 6-month period, we examined how the children interacted with VocSyl, their preferences, and most importantly, the challenges they faced when interacting with voice visualizations that targeted multisyllabic speech. Note that our goal was not to design one "perfect" visualization, but rather to design a robust suite of visualizations: the challenges and preferences of each child will vary. While user challenges and preferences are not the only factor that should be used to make final design decisions (see future work), the lessons learned from assessing design challenges, and understanding users has direct impact on improving usability and helping the broader HCI community working in this domain [45].

Participants

Three groups of children were recruited for the TCUID. Two children with ASD were recruited from a school for children with disabilities. Two children with SPD were recruited from the local community[9]. Four neurologically typical children [10] were recruited from a local preschool and child care program within the University of Illinois. All children were given an initial screening involving Part 1A of the *Communicative Development Inventory (CDI)* [12] and the Words Sequence section of the *Verbal Motor Production Assessment for Children* (VMPAC) [16]. Descriptive information about individual children across all groups is provided in Table 1. All children were past the age of 18-24 months during which multi-word productions typically emerge. The typically-developing children produced and understood all words on the CDI and correctly produced all syllables from the VMPAC, whereas the two groups with speech-language impairments demonstrated marked delays.

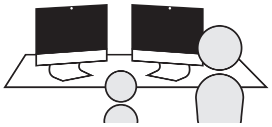


Figure 3. VocSyl TCUID Setup

The child would sit between the two computers while the researcher sat on the right side of the child. The child's chair would be a non-swivel chair.

Methods and TCUID Protocol

Our TCUID began with an initial speech-language screening performed by a faculty member or a masters degree student in Speech & Hearing Science. Once a child was deemed eligible according to the inclusion criteria, he/she participated in six study sessions. Study sessions were separated by approximately 7-14 days (based on the availability of the child and completion of VocSyl software modifications). Sessions lasted about 15 minutes and were conducted in an unused classroom at the child's school or in the Computer Science department at the University of Illinois.

Protocol

At the start of each session, the child would sit in front of two computers running VocSyl (Figure 3) and the researcher would explain the protocol in age appropriate vocabulary. Nowhere in the explanation did the researcher say, *"our visualizations should be touching."* This information was omitted to determine if the participants would find the graphical alignment of the visualizations to be naturally appealing.

Following the explanation, the researcher would then conduct four exposures. An *exposure* refers to a unique pairing of two visualization permutations (one on the left computer and one on the right computer). During the exposure, the researcher would use a simplified ABA [1] protocol. This process would begin with the researcher randomly choosing one computer and saying, *"Let's start over here."* The researcher would then say a target word (e.g., *"Monkey"*), while VocSyl visualized in real time. Then, the researcher would encourage the child to try (e.g., *"Now you say monkey"*). The child would then try to repeat the word while VocSyl drew their utterance. Regardless of the outcome (accurate or not), the researcher would provide verbal word encouragement (e.g., *"Good Try!"*). This process would be repeated using the second computers visualization. When both computers had the visualizations rendered, the researcher would ask the child, *"Which one did you like better?"* The selected visualization would be repeated. For Typical children, the researcher would also ask why they chose the particular visualization. Visualization order (which computer and visualization was used first) was randomized to minimize the impact of order effects. During each exposure, one word was used. Unlike a clinical setting where the goal is learning/impact, no words

[8] Soft BPF (gradually remove signal above/below a Hz threshold), hard BPS (sharp cut off signal above/below a Hz threshold), and High Pass Filter (remove signal below a Hz threshold)

[9] Given the comorbidity of ASD, SPD and other developmental disorders we cast a "wide net" with our recruitment to ensure rich and robust set of TCUID feedback from potential users.

[10] Neurologically typical children were recruited because they can verbally justify their preferences and articulate their confusion.

Name	Gender	Age	Diagnosis	CDI Word Count	VMPAC Syllable Count
Sean	M	4	SPD	231	7
Tina	F	8	SPD *(Smith-Magenis)*	308	3
Zev	M	10	ASD	23	0
Frank	M	14	ASD + Downs	198	7
Cara	F	5	Typical	396	13
Heather	F	6	Typical	396	13
Emma	F	5	Typical	396	13
Amy	F	6	Typical	396	13

Table 1. Demographics

Names were changed to protect identity, but gender was maintained. VPAC Syllable Count from Items #61-64

were repeated across exposures or sessions to minimize preference or other bias from learning.

Word Selection

Target words were selected to challenge childrens productive abilities and thus emulate the intended use of VocSyl. For the children in the ASD and SPD groups, target words were selected based on results from the CDI [12]. Specifically, the examiner chose multisyllabic words that the child reportedly "understands but does not say. The words selected were either two- or three-syllable words. Given that the children from the typical group could reportedly say all words listed on the CDI, three- or four-syllable nonsense [9] and low-frequency vocabulary words[28] were used instead. Overall, if children *could* imitate the words (even matching syllables or intonation) it suggests the potential applications of VocSyl within a speech therapy context.

Technology

The TCUID experiment utilized two 23" or 21" iMacs. For children with ASD or SPD, we used a Phoenix Duo USB Microphone with each computer. For typical children, we used the internal iMac microphones. The reason for the different microphones is discussed later in the paper. Sessions were videotaped using SilverBack software that records both the screen and live video (using the iMac iSight camera). All video was recorded with parental approval.

TCUID Iteration Results: Six Rounds

We will now present the observations from the initial speech-language assessment and all six rounds of our TCUID process. For each round we briefly describe what the round was "examining," followed by observations[11] of the typical children, then the ASD/SPD children. Round one examined the impact of visualization style (e.g., circle vs. envelope or solid circles vs. empty-circle). Rounds 2-6 each examined specific features of our visualizations by comparing a visualization with a feature turned "on" to the same visualization with the same feature turned "off." Decisions on which features to bring forward to future rounds were made based on feedback within *and* across all children (to account for broad and demographic specific feedback). Recall that we refer to the clinician's prompt/example of the word as a **model** and the child's response as the **attempt.**

While we do include the raw preference counts in Table 2, due to the small sample size, conducing traditional statistical comparisons is not done due to limited statistical power. These values are included for illustrative purposes only.

Initial Speech Assessment Observations

While the assessment session was not technically a "data gathering session," we noted that children with ASD and SPD were extremely quiet. They were willing to participate, to varying degrees, but their speech volume was low, and we realized the internal microphones on the iMacs would not pick up their vocalizations. We therefore modified our protocol to include the use of USB microphones that could be moved closer, or given to these children to hold based on the skills of each individual child.

Round 1 - Style

During the first round we examined the style of visualizations (see Table 2). For each exposure, we contrasted one visualization style with another (e.g., Figure 2A vs. 2C.).

Round 1 was the children's first exposure, and many seemed excited by the computer's response to their voice. Many of the typical children commented on the visualizations' shape or alignment (e.g., Emma said, *"Mine is like green toothpaste!"* and Amy said, *"They're matching up with mine!"*). This reaction was observed throughout the study.

When interacting with VocSyl, Zev often would lean into the iMac and touch the visualizations as they appeared on the screen. He clapped when the visualization appeared and said, "I did it!" during the session. Zev, Tina and Sean appeared very happy to be using the software and began smiling once we began using the software and it reacted to their voice. Tina often exclaimed, "Wow!" when the visualization appeared. In addition, Tina also held the microphone and put it up to her mouth *only* when it was her turn. She would continue this behavior throughout the entire six rounds.

Overall, this round produced no consistent preference across or within our participants. All the children actively participated. They attempted all the words presented to them. All appeared to enjoy themselves.

Round 2 - Split Screen vs. Layered Screen

During this round, we examined the physical relationship of the visualization of the clinician's model to the visualization of the child's attempt. One computer in each exposure would

[11] The qualitative observations and inferences in TCUID with children that cannot express their opinions are inherently subjective.

Round	A vs. B			A Count	B Count
1 *	Empty Circle	*vs*	Solid Circle	1	4
1 *	Envelope	*vs*	Solid Circle	4	2
1 *	Envelope	*vs*	Empty Circle	2	4
1 *	Circle+Line	*vs*	Circle Only	4	3
2 †	Split Screen	*vs*	Layered	12	19
3 †	no Pitch	*vs*	with Pitch	15	16
4	no DPA	*vs*	with DPA	16	16
5	no Images	*vs*	with Images	6	26
6 §	no Stoplight	*vs*	with Stoplight	16	12

Table 2. VocSyl Preferences
These values are included for illustrative purposes only.
∗Sean would not give preferences this round — § Tina was unable to participate in this round due to medical issues. — † During one exposure, one child did not give a preference (thus total preferences was 31 not 32)

randomly have the model visualized on the top of the screen with the attempt visualized on the bottom (Figure 2 D), while the other (with an identical visualization) would have the two visualizations overlaid (Figure 2 A, B or C).

Similar to Round 1, the verbal reactions of the typical children continued to be positive and analytical. These children regularly noticed patterns and relationships between the visualizations. Amy commented *"Boy, those ones are touching very good!"*, indicating that both the visualizations of the model and her attempt touching was a positive thing (despite not being mentioned in the protocol). While the reactions of Zev, Tina and Sean remained the same, Frank began to smile and readily engage with the researcher and VocSyl. While he was phonetically quite inaccurate, he matched the syllables and inflection of the researcher on each attempt. After his attempts, he would often smile as he looked at the computer screen. We did not see this response during Round 1.

The second round generally saw a preference for layered visualizations, however, there was a noteworthy observation. When we tested envelope split screen (Figure 2 D) versus envelope layered (Figure 2 C), all eight children preferred the envelope layered. However, for the other three exposures (which utilized variations on the circles), the preferences were split 50/50 between split screen and layered. This perfect split was also seen within the typical and ASD/SPD groups. Going forward, we opted to leave all visualizations layered due to the overall preference.

Round 3 - Pitch

It could be argued that with four features being displayed (volume, syllable, pitch and timing), this may be "too much" for the children to process. For Round 3, we reduced the complexity of the visualization by examining the role of pitch in the visualizations. Visualizations without pitch (no variation in the y-axis) would render the visualizations closer in presentation to the pacing board. One computer in each exposure would have pitch "on" (allowing the y-axis to be used), while the other (with an identical visualization) would have pitch turned "off" (each syllable's y-offset was 0).[12]

In Round 3, the children had a slight preference for having pitch on. When asked to explain their preference, the typical children most often cited the relative position or alignment of the model and the attempt as their justification. Amy justified her preference by saying, "Because I like how it goes down and the circles are touching." Emma gave a similar rationale stating she chose one over the other, "because we have the same number [of circles]." Zev's reaction to pitch was also positive: he smiled, leaned closer to the computer and clapped when the pitch-on visualization came up. Given the positive reaction, we decided to continue using pitch in the visualization for the remaining three rounds.

A secondary observation in Round 3 is that all the typical children (and Sean from the SPD group) began asking to choose the color of the visualization. Color is a variable controlled by researchers and was not initially intended to be an option for the children. However, seeing that color preference was important to some children, we obliged. From this round forward, children changed the visualization colors during each exposure (though the color was always the same on both computers within any given exposure).

Round 4 - Dynamic Production Alignment

Over the first 3 rounds, we noticed that all of the typical children would often justify their preference by commenting on how closely aligned the model and the attempt were. At times, however, the childrens' vocalizations would not directly align because their first syllable may have taken longer to say or their more explicit intake of breath would start the software. We hypothesized that many of the preference decisions were a result of the childs desire for alignment and matching the model. This would, in theory, be a positive observation in behavior in that it is the main goal of VocSyl.

To further explore the role of alignment in VocSyl, we created a feature called **D**ynamic **P**roduction **A**lignment (**DPA**). When DPA is activated, after the child has made their attempt, their vocalization would slide to the left or to the right in one or two seconds such that their first syllable would be directly on top of the researcher's first syllable. This would align the two productions, providing a more direct comparison between the model and the attempt. One computer in each exposure had DPA turned on, while the other (with an identical visualization) would have DPA turned off.

Seven of the eight children took note of the movement in the DPA (either by verbally commenting, or visibly reacting to the animation). Notably from the ASD/SPD group, Frank exclaimed "I like!" when he first saw the DPA, and Sean turned to his mother twice, saying, "it goes ma ma," and "it's moving mommy." While preference was split (50/50), given the observed positive reaction and increased attention to VocSyl, we elected to enable DPA in Rounds 5 and 6.

It is worth noting that in Round 4, Tina explicitly asked (with very limited articulation), "Do more," after we finished with the required exposures. So as not to bias her to future VocSyl exposures, we turned on an oscilloscope visualization, which we had built previously for debugging VocSyl. Tina subsequently played with the oscilloscope twice, each lasting for about a minute.

[12] Pitch *was* turned on during Rounds 1 and 2.

Round 5 - Found Images

Images related to the childs interest are likely to spur engagement [14, 34]. During this round, we examined the impact of replacing the abstract circle with found images (Figure 2B replaces the attempt's circles with trucks). One computer would randomly replace the circles of the child's attempt with a found image, while the other (with an identical visualization) would use the original circles. *As there are no objects that could naturally replace the envelope visualization, we did not use it as one of the four exposures.*

We asked each typical child for a cartoon they liked, and parents/caregivers of the ASD/SPD children for preferences or interests. Overwhelmingly, children's preference was to have the images in VocSyl over the abstract shapes. Preference from the typical children was based on the presence of images (e.g., "I want Cinderella again! .. my path is Cinderella!" - Emma, "because Elmo, I like Elmo" - Cara). Preferences for the ASD/SPD children were equally based on the images. Tina attempted to count and point at the found images, Frank said "ohhhhhhh" with a look of immense pleasure on his face, while Zev would say "truck."

We observed that when children saw researchers turning on and off the images, both typical and SPD children started asking or pointing to the screen to choose their own images. Similar to the color selection that occurred in Round 3, an option to change the image was not intended by researchers, but was identified by the children.

Round 6 - Stoplight Cue

We noted that children had difficulty waiting to start their attempt. A few times, the children were so excited to say their word and see the visualization, that as soon as the researcher finished saying the model, they did not wait for VocSyl to start rendering their production (a process that takes seconds). We designed a visual cue to show when it is the child's turn. Specifically, after the researcher finished the model, a stoplight appears and changes from red to yellow to green (over three seconds). When the stoplight rests on green, VocSyl begins listening to the attempt by the child.

We had a negative reaction from all children. When asked for a justification, we were told that they disliked waiting (e.g., "because I dont like to wait" - Amy, "when I do it with the stoplight, I have to wait" - Emma). Frank, rather than waiting, continued to repeat his attempt until the computer executed his visualization. This explicit delayed interaction and gratification seemed to be less desirable, even though the children now knew exactly when to begin to interact.

In addition, Sean, Zev, and two of the four typical children asked (verbally or by pointing) to have the found images turned on in this round. Also, Frank would move his hand horizontally in time with the DPA aligning animation. In addition, Frank spent almost a minute exploring his voice with the oscilloscope after the conclusion of the study.

General Observations

We observed that all the typical children commented or narrated the visualization. Comments centered on the shape of the visualizations, who had "more circles," and when or where the visualizations touched. Cara, in particular, personified the circles as family members. She would label the "mommy" circle, the "daddy" circle, and the "baby" circle (based on size). Children with ASD and SPD, however, were not generally explicit or commented on the visualizations. This may be a direct result of their linguistic and/or developmental challenges. Even though this was not a treatment study, it is worth noting that the children with SPD and ASD were actively attempting words that they do not say. More impressively, both children with SPD were saying the words well. The children with ASD made good attempts, almost always matching the syllables and pitch (and Frank regularly matching phonemes). Occasionally, Frank and Sean would repeat the same phonetic pattern across sessions (regardless of the model). We believe that this may be tied to words they found challenging, and rather than failing, they wanted to see a visualization react to their voice.

The last observation returns to the observed issues with "volume" during the initial speech assessment. While Tina, Sean and Frank were quiet during their initial assessment and early interactions with VocSyl, over the TCUID process they became increasingly loud with interacting with the computer. By Round 6, Sean was yelling his words into the computer, seemingly to get the biggest response possible.

Positive and Enriching Experience

All the children appeared to enjoy engaging with the computer and wanted to continue their interactions. We explicitly asked the children from the typical group whether they had fun and if they wanted to come back. Tina and Sean both asked to use the computer again (though these requests were "translated" by their parents). Frank and Zev often talked into the microphone after exposures finished, continuing to look at the screen for a visualization of their voice.

This engagement is a important element of a positive learning environment, in that engagement is tied to learning [39]. Over the six rounds, the children became increasingly relaxed and engaged with VocSyl. All children (including the children with ASD/SPD) came into sessions smiling, rarely needing a prompt to sit down at the computers. Further, all the children attempted the words they were prompted with. This is particularly meaningful considering that all the words tested with the four children with ASD or SPD were words that they were not reported to use prior to starting the study. Though the phonetic accuracy varied by each child, attempts were made by all. Syllables and intonation were almost always correct. Tina's mother particularly commented on her daughter's ability to say the study's words, remarking how Tina never says these words at home *and* she was not currently enrolled in any speech therapy. Tina's mother also asked for DVDs of her daughter's session to re-watch her daughter saying all the new vocabulary.

Physical Interactions

At the conclusion of each session, we allowed the children to play with an iPad. This was not an interaction to be explicitly tested as part of the TCUID, but rather a reward for participating. During our sessions, we noticed that the children (especially the ASD and SPD children) readily enjoyed playing with the iPad and smiled when they got the applications to respond to their touch. Based on this, we re-watched the recordings of the six rounds of data collection. During this, we noticed that all typical children regularly touched the screen of the iMacs to describe the visualizations they saw and Sean, Frank and Zev both attempted to interact with the visualizations on the screens of the iMacs.

DESIGN GUIDELINES

From our observations of the TCUID, discussions with typical children and the existing literature, we have produced a set of guidelines for designing software to facilitate multisyllabic speech production. While some guidelines may appear "obvious," this work grounds these design considerations in experimental observation, rather than anecdotal knowledge.

R1 Minimize Delay to Interaction: When designing software for children, ensure that when the child wants to engage, and the software is ready to respond and delays are minimized. Thus children are engaged and stay interested with the interaction, the software, and learning.

As we saw with the stoplight condition, and the children's desire to keep interacting, the computer visualizations are a highly motivating media for vocalization. However, the more delays we build into the system, the more frustrating and confusing these interactions appear to be.

R2 Real-Time is Fun: By showing visualizations change in real time, children's attention remains with the software. They then continue to perform the task/activity. By ensuring real-time visualizations, we hopefully can encourage learning [39].

Given that our tasks were well-structured, we were unsure if they would still encourage interaction (as compared to the free-form task of [14]). Our observations appear to suggest that children truly enjoyed playing with the computer, as their requests to keep going, verbal or otherwise, continued even after the official session had concluded. As visualizations change and animated in real-time, children both discussed the changes and moved in concert with the animations. This avenue of research is promising, and supports further exploration of how complex interactions can get while still encouraging skill development and engagement

R3 Child Customization: Interaction designers should support simple, uncluttered option menus for children to make choices in the interface themselves – ideally allow children to point to and touch what they want. This ensures that as the child's own preferences change (within and between sessions), so can the software. By designing easy to use interfaces, children can feel empowered and engaged with their interactions. By employing touch, non-verbal children can customize the software themselves.

The degree of customization children requested with VocSyl was an unexpected finding. Both typical and ASD/SPD children requested color changes or picture changes, as soon as they observed that the researcher was changing a setting (or with color, that there was a grid of colors on the settings panel). This was a natural and unprompted gesture seen in nearly every instance that the children greatly enjoyed. While the work of Morris [34] suggests that customization can be facilitated in higher functioning children, it is noteworthy that children here actively sought out customization in the VocSyl interface.

R4 Dynamic Computer Correction: When the target child is learning a complex skill (e.g., speech), the software can "smooth" their interactions by auto-correcting or auto-adjusting minor mistakes without negatively impacting child's interaction. This can allow the clinician to focus on the skills being targeted rather than trying to explain minor errors by the child (or the software).

We explicitly manipulated word attempts by children using DPA. Given the positive reaction to DPA, we believe that systems, such as VocSyl, need not be completely literal in their visualization. Computers provide a unique ability over non-dynamic tools (Figure 1) to correct tiny mistakes made by a child or a researcher in order to focus on the goal therapy. It appears that if changes made are obvious to the child (e.g., sliding of the production rather than a sudden jump), they are accepted and understood.

R5 Robust Microphone Setup: Systems should provide a robust setup to accommodate multiple voice levels, as children have different comfort levels with the use of their voice. As children are less sure of their abilities, they may be more hesitant to loudly engage. As their comfort level rises, so will their voice level.

We found that neurologically typical children were more willing to speak loudly than children with ASD/SPD. Internal computer microphones did not suffice. Even an external USB microphone may need to be placed very close to the child. A USB microphone also provides a physical device that the child can hold onto and direct their engagement towards.

R6 Competence of the Child: Designers can "raise the bar" on targeted tasks and make them more challenging by allowing the clinicians to adjust the "picky-ness" of the software for assessing correctness. Further, tasks asked of children should likewise be able to become more complex.

While one half of the children in our study had speech delays, they were capable of successful interactions with VocSyl, regardless of complexity. We therefore encourage designers to be cautious with respect to oversimplification of software. Take on fairly complicated projects in this direction, with the option to dial back features. Further, given the interactions observed with Frank and Zev we strongly believe that all the children fundamentally understood that their voice created and manipulated the visualization.

R7 Physical Interaction: Children want to touch everything. Touch is an easy-to-understand interaction. Therefore, design systems that not only respond to touch, but provide meaningful feedback for those interactions.

We observed physical interaction with the computer on multiple occasions. We believe that large screens, animations, and vibrant shapes prove an engaging visual experience. All but two of the children attempted to touch the computer screen to interact with the shapes or to comment on them. We strongly believe that encouraging physical interaction can have an impact on the visualization and can provide therapists with new techniques to further teach and shape vocalization.

VOCSYL TOUCH

Following our unexpected observations about physical interactions with the iMacs, we have ported the VocSyl system to the iPad and iPhone with the potential for easy distribution to homes and clinicians. Most notable to this version was the addition of touch interactions. While the VocSyl Touch implementation visually looks and reacts the same as the desktop version, when a clinician or child touches the syllables in the visualization, VocSyl slightly animates that touch and plays back that segment of the production from a live audio recording.

FUTURE WORK AND LIMITATIONS

Given the small scale and qualitative nature of this study, we cannot conclude that VocSyl will explicitly improve speech therapy and multisyllabic speech production (though qualitative observations do appear promising). However, this was not the aim of the current research. Prior to running a fully powered study, we aimed to ensure that the VocSyl and its interactions were designed carefully, with consideration for the end-user. We are currently conducting a mixed method intervention study of 18 children with speech-language impairments enrolled in one of three conditions: intervention with VocSyl, traditional therapy with a pacing board, and a playgroup aimed at social interaction. The goal is to examine the impact of all three conditions on children's multisyllabic productions. This intervention study will also allow us to explore the impact of these features on learning. Further, as we continue research on VocSyl, incorporating more rigorous qualitative and quantitative analysis (computational and hand-coded[15]) is key.

One additional avenue of exploration with VocSyl is phoneme production. Some of the children in our study have speech sound (phoneme) impairments. We believe that the system could be expanded easily to teach and shape phoneme acquisition, and are actively expanding VocSyl to target this. Alternatively, a future experiment could explore the impact of algorithm errors (e.g., when syllables are misdetected) on language acquisition.

We also wish to highlight that there is a leap between producing multisyllabic speech and real-world communication. Our study focused on encouraging a specific behavior that is one component of functional communication. This work, in conjunction with the findings from other research, lays the groundwork for future exploration of this area.

CONCLUSION

The primary aim of this research was to design a software tool called VocSyl for use in speech therapy to encourage multisyllabic speech production in children with ASD and SPD. Our goal was to include children with ASD and SPD *in* the design process through TCUID so as to emphasize building what the intended users demonstrated they wanted. This user-centered approach allowed us to design a software system that is both highly configurable and provides a platform for a meaningful exploration of multisyllabic speech productions. In addition to developing our software system, we also generated design guidelines for future research and software working with this population to teach speech skills.

Given the overwhelmingly positive response from children and parents to VocSyl, we believe that our software and computer-based interventions have the potential to improve multisyllabic speech production in children with ASD and SPD. While this was *not* a treatment study, the children involved were actively saying words using VocSyl that they were not saying at home. This adds further support to the potential impact of computer-based visualizations in speech therapy.

ACKNOWLEDGEMENTS

We would like to thank all the children and their families for participating in this project. Also the Autism Speaks Dennis Weatherstone Fellowship (6312), Autism Speaks (5744) and NSF (10-29035) for the funding and support of this research.

REFERENCES

[1] Baer, D., Wolf, M., and Risley, T. Some current dimensions of applied behavior analysis. *Journal of applied behavior analysis 1*, 1 (1968), 91–97.

[2] Barry, R. M. Epg from square one: An overview of electropalatography as an aid to therapy. *Clinical Linguistics & Phonetics 3*, 1 (1989), 81–91.

[3] Baskett, C. B. *The effect of live interactive video on the communicative behavior in children with autism*. PhD thesis, UNC - Chapel Hill, 1996.

[4] Bondy, A., and Frost, L. The picture exchange communication system. *Behavior Modification 25*, 5 (2001), 725–744.

[5] Bryson, S. Brief report: Epidemiology of autism. *Journal of Autism and Developmental Disorders 26*, 2 (1996), 165–167.

[6] Center for Disease Control and Prevention, CDC. Autism information center, Date April 25, 2007.

[7] Consolvo, S., Everitt, K., Smith, I., and Landay, J. Design requirements for technologies that encourage physical activity. In *Proceedings of the SIGCHI*, ACM (2006), 457–466.

[8] Dettmer, S., Simpson, R., Myles, B., and Ganz, J. The use of visual supports to facilitate transitions of students with autism. *Focus on Autism and Other Developmental Disabilities 15*, 3 (2000), 163.

[9] Dollaghan, C., and Campbell, T. Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research 41*, 5 (1998), 1136.

[10] Edwards, W., Bellotti, V., Dey, A., and Newman, M. The challenges of user-centered design and evaluation for infrastructure. In *Proceedings of the SIGCHI*, ACM (2003), 297–304.

[11] Fell, H., MacAuslan, J., Gong, J., Cress, C., and Salvo, T. visibabble for pre-speech feedback. In *Extended abstracts of CHI'06*, ACM (2006), 767–772.

[12] Fenson, L., Dale, P., Reznick, J., Thal, D., Bates, E., Hartung, J., Pethick, S., Reilly, J., et al. *MacArthur Communicative Development Inventories: User's guide and technical manual*. Paul H. Brookes Baltimore, MD, 2002.

[13] Gamma, E. *Design patterns: elements of reusable object-oriented software*. Addison-Wesley Professional, 1995.

[14] Hailpern, J., Karahalios, K., and Halle, J. Creating a spoken impact: encouraging vocalization through audio visual feedback in children with asd. In *CHI 2009*, ACM (Boston, MA, 2009).

[15] Hailpern, J., Karahalios, K., Halle, J., Dethorne, L., and Coletto, M.-K. A3: Hci coding guideline for research using video annotation to assess behavior of nonverbal subjects with computer-based intervention. *ACM Trans. Access. Comput. 2* (June 2009).

[16] Hayden, D. *VMPAC: Verbal motor production assessment for children*. Psychological Corporation, 1999.

[17] Highman, C., Hennessey, N., Sherwood, M., and Leitão, S. Retrospective parent report of early vocal behaviours in children with suspected childhood apraxia of speech (scas). *Child Language Teaching and Therapy 24*, 3 (2008), 285.

[18] Hoque, M., Lane, J., El Kaliouby, R., Goodwin, M., and Picard, R. Exploring speech therapy games with children on the autism spectrum. In *Proceedings of InterSpeech*, Citeseer (2009).

[19] IBM. Speech Viewer III, Date 1997.

[20] Kanner, L. Autistic disturbances of affective contact. In *Nervous Child 2*, L. Kanner, Ed. V.H. Winston, 1943, 217–250.

[21] KayPentax. Visi-Pitch IV, Model 2950B, 1996-2008.

[22] Kerr, S. J., Neale, H. R., and Cobb, S. V. G. Virtual environments for social skills training: the importance of scaffolding in practice. In *Proceedings of ASSETS'02*, ACM Press (Edinburgh, Scotland, 2002).

[23] Kientz, J. A., Arriaga, R. I., Chetty, M., Hayes, G. R., Richardson, J., Patel, S. N., and Abowd, G. D. Grow and know: understanding record-keeping needs for tracking the development of young children. In *Proceedings of SIGCHI*, ACM Press (San Jose, California, USA, 2007).

[24] Leahu, L., Sengers, P., and Mateas, M. Interactionist ai and the promise of ubicomp, or, how to put your box in the world without putting the world in your box. In *Proceedings of the 10th international conference on Ubiquitous computing*, ACM (2008), 134–143.

[25] Lewis, C., and Rieman, J. Task-centered user interface design. *Available via ftp. cs. colorado. edu/pub/cs/distribs/clewis/HCI-Design-Books* (1993).

[26] Lovaas, I. I. *The Autistic Child*. John Wiley & Sons, Inc, New York, 1977.

[27] MacWhinney, B. Models of the emergence of language. *Annual review of psychology 49*, 1 (1998), 199–227.

[28] Mahurin-Smith, J. *The impact of prematurity on language skills at school age*. PhD thesis, University of Illinois at Urbana Champaign, 2010.

[29] Mateas, M. Expressive ai: A hybrid art and science practice. *Leonardo 34*, 2 (2001), 147–153.

[30] Mermelstein, P. Automatic segmentation of speech into syllabic units. *J. Acoust. Soc. Am 58*, 4 (1975), 880–883.

[31] Michaud, F., and Theberge-Turmel, C. Mobile robotic toys and autism. In *Socially Intelligent Agents - Creating Relationships with Computers and Robots*, K. Dautenhahn, Ed., vol. 3. Springer, 2002, 125–132.

[32] Millar, D., Light, J., and Schlosser, R. The impact of augmentative and alternative communication intervention on the speech production of individuals with developmental disabilities: a research review. *Journal of Speech, Language, and Hearing Research 49*, 2 (2006), 248.

[33] Minshew, N. J., Goldstein, G., and Siegel, D. J. Neuropsychologic functioning in autism: Profile of a complex information processing disorder. *Journal of the International Neuropsychological Society 3*. (1997), 303–316.

[34] Morris, R., Kirschbaum, C., and Picard, R. Broadening accessibility through special interests: a new approach for software customization. In *Proceedings of ASSETS'10*, ACM (2010), 171–178.

[35] Nakagawa, S. A survey on automatic speech recognition. *IEICE TRANSACTIONS on Information and Systems E85-D*, 3 (2002), 465–486.

[36] Preston, J., and Edwards, M. Phonological processing skills of adolescents with residual speech sound errors. *Language, speech, and hearing services in schools 38*, 4 (2007), 297.

[37] Roads, C., Strawn, J., Abbott, C., Gordon, J., and Greenspun, P. *The computer music tutorial*, vol. 81. MIT press Cambridge, Massachusetts, 1996.

[38] Shuster, L. I., and Ruscello, D. M. Evoking [r] using visual feedback. *American Journal of Speech-Language Pathology 1*, May (1992), 29–34.

[39] Skinner, E., Wellborn, J., and Connell, J. What it takes to do well in school and whether i've got it: The role of perceived control in children's engagement and school achievement. *Journal of Educational Psychology 82*, 1 (1990), 22–32.

[40] Strand, E., Stoeckel, R., and Baas, B. Treatment of severe childhood apraxia of speech: A treatment efficacy study. *Journal of Medical Speech Language Pathology 14*, 4 (2006), 297.

[41] Strik, H., and Cucchiarini, C. Modeling pronunciation variation for asr: A survey of the literature. *Speech Communication 29*, 2-4 (1999), 225–246.

[42] Tager-Flusberg, H., Rogers, S., Cooper, J., Landa, R., Lord, C., Paul, R., Rice, M., Stoel-Gammon, C., Wetherby, A., and Yoder, P. Defining spoken language benchmarks and selecting measures of expressive language development for young children with autism spectrum disorders. *Journal of Speech, Language, and Hearing Research 52*, 3 (2009), 643.

[43] Tartaro, A., and Cassell, J. Playing with virtual peers: Bootstrapping contingent discourse in children with autism. In *Proceedings of International Conference of the Learning Sciences*, ACM Press (Utrecht, Netherlands, 2008).

[44] Velleman, S. Phonotactic therapy. In *Seminars in Speech and Language*, vol. 23 (2002), 35–46.

[45] Vredenburg, K., Mao, J., Smith, P., and Carey, T. A survey of user-centered design practice. In *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*, ACM (2002), 471–478.