# Voicesetting: Voice Authoring UIs for Improved Expressivity in Augmentative Communication

**Alexander J. Fiannaca, Ann Paradiso, Jon Campbell, Meredith Ringel Morris**
Microsoft Research, Redmond, WA, USA
{alfianna, annpar, joncamp, merrie}@microsoft.com

## ABSTRACT
Alternative and augmentative communication (AAC) systems used by people with speech disabilities rely on text-to-speech (TTS) engines for synthesizing speech. Advances in TTS systems allowing for the rendering of speech with a range of emotions have yet to be incorporated into AAC systems, leaving AAC users with speech that is mostly devoid of emotion and expressivity. In this work, we describe *voicesetting* as the process of authoring the speech properties of text. We present the design and evaluation of two voicesetting user interfaces: the *Expressive Keyboard*, designed for rapid addition of expressivity to speech, and the *Voicesetting Editor*, designed for more careful crafting of the way text should be spoken. We evaluated the perceived output quality, requisite effort, and usability of both interfaces; the concept of voicesetting and our interfaces were highly valued by end-users as an enhancement to communication quality. We close by discussing design insights from our evaluations.

### ACM Classification Keywords
K.4.2 Social Issues: Assistive Technology.

### Author Keywords
AAC; TTS; Amyotrophic Lateral Sclerosis; ALS.

## INTRODUCTION
Augmentative and alternative communication (AAC) devices are technologies that allow people with speech disabilities to communicate. Current speech generating devices (SGD) used for AAC offer little in the way of allowing users to control the expressive nature of the speech rendered from the user's input [11]. This is surprising given that advances in speech synthesis technologies over the past decade have resulted in the development of text-to-speech (TTS) engines capable of rendering speech that exhibits a range of emotions (e.g., CereProc [4], Nuance Loquendo [18], and IBM Watson [30] are commercial speech engines capable of emotional speech synthesis). Additionally, most TTS engines accept Speech Synthesis Markup Language (SSML) [2] as input, allowing for a degree of control over prosodic features such as the rate of speech, the pitch of the voice, and cadence/pacing of words. The fact that these advances have yet to be incorporated into SGDs in a meaningful way is a major issue for AAC. Recent work such as that of Higginbotham [9], Kane et. al. [11], and Pullin et al. [22] has described the importance of this issue and the need to develop better AAC systems capable of more expressive speech, but to date, there are no research or commercially available AAC devices that provide advanced expressive speech capabilities, with a majority only allowing for basic modification of speech parameters (overall rate and volume) that cannot be varied on-the-fly (as utterances are constructed and played), while not leveraging the capabilities available in modern TTS engines.

We address this issue by designing voice authoring interfaces that apply state-of-the-art TTS engine features in order to create more expressive speech from AAC devices. The problem our work addresses is not whether AAC systems can automatically generate more expressive speech, but rather how AAC systems can be designed to allow the user to control speech expressivity themselves. We introduce the concept of *voicesetting,* the process of authoring/editing the speech properties of text (analogous to typesetting for the visual display of text), and contribute two novel interface designs to support voicesetting by AAC users. *The Expressive Keyboard* allows for rapid expressivity through the insertion of emoji and punctuation into text. Additionally, the Expressive Keyboard includes an *Active Listening Mode* that allows users to rapidly respond while listening to others speak by playing expressive vocal sound effects like laughter or scoffing. *The Voicesetting Editor* trades off the low cost of use in the Expressive Keyboard for a higher degree of control over the exact speech properties of the output. Finally, we present an evaluation of these two interfaces and a discussion of extensions to this work that could further increase expressivity.

## RELATED WORK

### Expressivity in Current AAC Devices
In an interview study with people with Amyotrophic Lateral Sclerosis (ALS) (a degenerative neuromuscular disease often necessitating use of AAC), Kane et. al. [11] found that many AAC users are dissatisfied with the expressivity and tone of the synthesized voice they are able to generate through AAC devices. Conveying emotions while speaking uses non-verbal communicative features such as gesture and prosody [23], however, current AAC devices do not support the


*CHI 2018*, April 21–26, 2018, Montreal, QC, Canada

ACM 978-1-4503-5620-6/18/04…$15.00
https://doi.org/10.1145/3173574.3173857.

conveyance of this non-verbal information, and generally only provide text-to-speech engines and voice fonts capable of synthesizing speech with a single, flat tone. In the Kane et. al. study [11], participants indicated that they often type and speak additional explanatory phrases such as "I am angry" before typing and speaking their intended phrase. Given the slow rate of text entry in many AAC devices (10 to 20 words per minute [13]), explicitly typing expressivity statements presents a significant effort (in addition to being unnatural). In a design investigation with children, Light et. al. [12] noted the lack of expressive features.

Portnuff [21] proposed a hierarchy of needs for the design of speech-generating AAC systems, with five of the nine levels dealing with expressivity ("pitch & speed control," "expressiveness," "multilingual capability," "loudness," and "ability to sing"). Most AAC devices currently only support the bottom two levels of Portnuff's hierarchy ("Windows OS compatibility" and "intelligibility"). From the Disability and Rehabilitation perspective, Nathanson [17] described the effects of the substitution of the voice a user is losing (in the case of acquired disabilities like ALS) with these inexpressive AAC systems as having "profoundly negative impacts" on self-concept and identity, making the case that the current lack of expressive AAC systems is a moral and ethical issue. Higginbotham noted that the limitations in the expressivity of voices on AAC systems has resulted in the external synthesized voice not representing the internal voice (e.g., self-identity) of their users [9].

### Expressivity in Text-to-Speech Engines

While AAC devices are significantly lacking in the ability to generate expressive speech, a large amount of speech synthesis research has aimed at creating more expressive speech engines [24]. Synthesizing human-quality expressivity in synthetic speech is still an open problem, but significant advances have been made. Speech engines such as Pitrelli et. al.'s [20] have the ability to render speech that listeners can identify as exhibiting several different tones (e.g., "conveying bad news" or "asking yes-no questions"). Several commercial engines such as the IBM Watson [28], CereVoice [4], and Nuance Loquendo [18] allow for the rendering of emotional synthesized speech. Each of these systems is similar in allowing developers to select between a small set of emotional categories when rendering speech (e.g., CereVoice supports a default voice in addition to "happy," "sad," "calm," and "cross"). Recent developments on the MaryTTS engine (an open-source TTS research platform) [14] have extended beyond simply accepting the specification of emotional categories, allowing for emotional tones to be specified with continuous values for emotional dimensions such as pleasure, arousal, and dominance [5]. In addition to allowing for the high-level specification of a tone of voice, each of these TTS engines also supports low-level speech modifications through the Speech Synthesis Markup Language (SSML) [2]. SSML is an XML-based language for specifying prosodic information such as the rate of speech, volume, baseline pitch, emphasized words, and pauses.

### Explorations into Expressive Speech Systems

Several recent projects have attempted to either create expressive AAC or explore the design space of expressive AAC. Sobel et. al. [25] articulated a design space for partner-facing visual displays ("awareness displays") for improving the expressivity of AAC by presenting visual feedback such as emoji, colored light patterns, text, or animated avatars to communication partners to augment synthesized speech. WinkTalk [27] is a prototype of a communication system that tracks a user's facial expressions to decide which of three expressive voices should be used to render synthetic speech for a particular utterance; however, a user study found that users prefer manual selection of the expressive features of their text over automatic selection via facial expressions.

Using Critical Design, Pullin and Hennig [22] describe three designs intended to provoke discussion and encourage thought about the complexities of tone of voice and its importance in AAC. In this work, the authors propose that AAC users should be able to author tones of voice and save and share these tones with other AAC users. The tone authoring process was explored through the ToneTable art piece, which accepts speech audio as input and allows able-bodied users to physically manipulate the tone of that speech [1]. While ToneTable provoked critical thought around the expressivity of speech, it was not a system designed for AAC, or intended to be used by people with speech impairments. Pauletto et. al. [19] investigated the design space of integrating emotional TTS into conversational bot systems. They proposed a brief set of possible design approaches for creating an interface for marking up text with speech attributes, though these design suggestions were not intended for use by AAC device users (i.e., design suggestions involving direct manipulation likely will not work well for an eye-gaze user). They also describe possible designs for displaying speech properties in text.

In this work, we connect the advances made in speech synthesis technology with the need for expressive AAC by designing, implementing, and evaluating interfaces that allow users to author the expressivity of their synthetic voice. This work presents a novel contribution to the field of AAC user interfaces, as no commercial or research AAC systems provide simple interfaces for prosodic manipulation and fast, non-speech backchannelling.

## SYSTEM DESIGN

The lack of expressivity in synthesized speech through AAC is a critical problem that has yet to be addressed, but the technology required to add expressivity to speech already exists in modern TTS engines. Therefore, we designed several voice authoring interfaces to allow AAC users to add expressivity to their speech. Our interfaces were built as an extension to a generic gaze-based on-screen keyboard designed to run on a Microsoft Surface Pro 3 tablet with a Tobii EyeX eye-gaze sensor. (Gaze-based AAC systems are typically used by people with severe motor disabilities, such as people who have advanced ALS or who are paralyzed.)

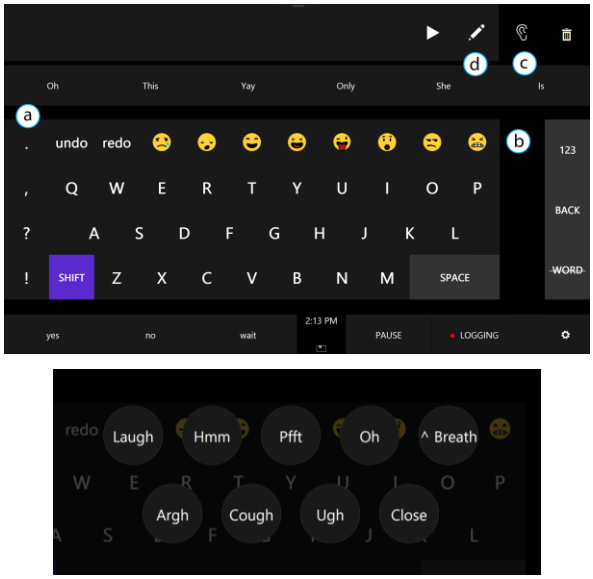


**Figure 1. Emoji (top, b) and punctuation marks (top, a) act as operators over text typed in the Expressive Keyboard (top), changing the synthetic speech output. The ALM (bottom) is initiated by selecting the "ear" key in the Expressive Keyboard (c); it is displayed as a set of buttons overlaid on the keyboard, each corresponding to a different vocal sound effect.**

In order to generate expressive synthetic speech, our system uses the CereVoice TTS engine [4]. This engine supports SSML with two expressive extensions. The first allows for specifying emotional tone (including happy, sad, calm, or cross/angry), and the second allows for the insertion of vocal sound effects like laughter or scoffing into the synthesized speech. In order to provide SSML input to the CereVoice TTS engine from our on-screen keyboard, we built a backend system that tracks input text and speech properties set over regions of that text, allowing for the generation of SSML. In the following sections, we describe the design of two novel interfaces, the Expressive Keyboard and the Voicesetting Editor, which allow users to provide the input that our backend system uses to generate the expressive output.

We employed an iterative design process wherein a group of people with ALS (PALS) provided informal feedback on the shortcomings of their current AAC systems and reactions to our proposed voicesetting concepts, thereby helping shape our interfaces to have functionality usable by and useful to PALS. We also conducted usability testing with able-bodied participants to identify and rectify usability concerns with the interfaces. For brevity, we present only the final designs.

### The Expressive Keyboard

Given that gaze-based text entry methods are currently extremely slow and laborious (typically 10 to 20 words per minute [13]), it was important in the design of our voicesetting interfaces that users be able to add expressivity to their face-to-face conversations without adding a large burden of additional time and effort. Therefore, our first interface, the Expressive Keyboard, was designed to allow users to specify the expressive nature of their speech by only adding one or two additional characters to the text they input. The Expressive Keyboard extends a standard gaze-based on-screen keyboard with a set of emoji and punctuation that, when inserted, act as operators over the surrounding text, changing the nature of the speech that will be generated from the text. Emoji and punctuation carry out different classes of operations (Figure 1, top).

| Name | Emoji | Voice | Vocal Sound Effect |
|---|---|---|---|
| Sad | | Sad | |
| Calm | | Calm | |
| Happy | | Happy | |
| Funny | | Happy | Laugh |
| Sarcastic | | Happy | Sarcastic Scoff |
| Surprised | | | Sharp Breath In |
| Irritated | | Cross (Angry) | |
| Angry | | Cross (Angry) | Argh |

**Table 1. Operations associated with emoji in the Expressive Keyboard. Vocal sound effects are always added at the beginning of the synthesized speech, and voices always span the entire sentence in which the emoji is present.**

Emoji affect the tone and emotion of the output speech at the sentence level. Emoji can change the tone of voice from the default tone to any of the emotional states provided by the CereVoice engine: "happy," "sad," "calm," and "cross". We bind this effect to the sentence in which the emoji is present. Additionally, emoji can insert vocal sound effects at the beginning of the sentence in which the emoji is present. Vocal sound effects include utterances like laughter or scoffing. Table 1 shows the emoji in the Expressive Keyboard and the operation associated with each. Our selection of emoji was influenced both by Eckman et. al.'s [6] work on culturally universal emotions and the description of emotions ALS patients most want to express according to interviews by Kane et. al. [11]. While we could add many more emoji to the system, we designed the current set to strike a balance between covering a broad range of emotions and restricting the number of options to keep the interface simple and usable through eye-gaze.

Punctuation carry out operations on a local level, changing prosodic features of their surrounding words. Periods, commas, and exclamation points insert a customizable amount of silent space between pronouncing words or sentences, allowing users to set the cadence of their speech. Adding two consecutive question marks will raise the pitch at the end of the sentence in order to emphasize the fact that a sentence is a question. This can be useful in scenarios in which the user asks a question that could have been a statement if not for the question mark (e.g., "She is meeting us there." vs. "She is meeting us there?").

#### *Active Listening Mode (ALM)*

While testing the Expressive Keyboard, it became apparent that using emoji without text in order to play only the vocal

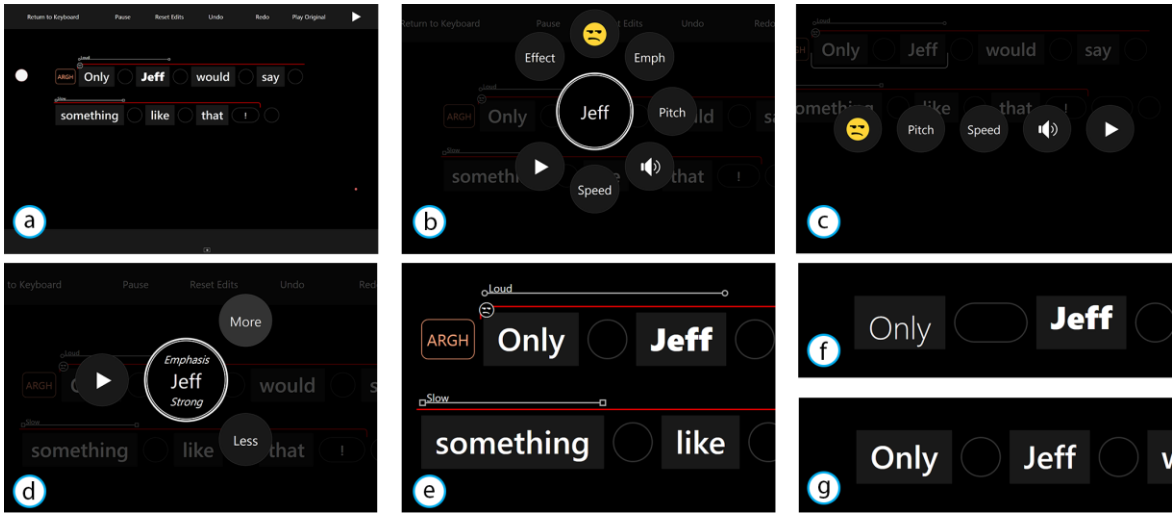


**Figure 2. The Voicesetting Editor interface parses and displays input text as word, punctuation, and vocal sound effect tokens (a, e). Speech property values can be edited on a single token basis (b) or over a range of tokens (c). Speech property values are only allowed to take on values from a set of predefined levels with semantically meaningful names (d). Speech property values are displayed using visual properties in the editor (e, f, g). Emotional tone, volume, and rate of speech are displayed as lines over token buttons with value descriptions like "Loud" at the outset of the line (e). Voice pitch is displayed using baseline height (f, g - higher baseline height corresponds to higher pitch). Emphasis is displayed using font boldness (f, g – bolder font corresponds to heavier emphasis). Oval width corresponds to pause length associated with inter-word or punctuation tokens (f depicts a longer pause than g).**

sound effects associated with those emoji could be useful while users are listening to communication partners speak. This interaction was akin to the gestures (e.g., nodding) [7] and non-verbal vocalizations (e.g., laughing) used in standard communication to communicate information from a listener to a speaker (in linguistics, this secondary channel of communication from a listener back to a speaker is known as *backchannel communication* [30]). This led us to develop the Active Listening Mode (ALM) of the Expressive Keyboard (Figure 1, bottom). The ALM provides single-click access to a variety of vocal sound effects. Where AAC users would typically have to type out statements (e.g., "I'm listening," "That's funny") in order to interact with an active speaker, the ALM allows listeners to rapidly express reactions to active speakers in only the time it takes to dwell-click a single button. The ALM currently provides the following vocal sound effects: laughter, "hmm," a sarcastic scoff ("pfft"), "oh?," a sharp intake of breath, an angry "argh," a cough, and a disgusted "ugh." Each of these reactions is currently generated by the CereProc TTS engine, although conceivably these could be customizable by end users, allowing them to use any sound files or even their own voice-banked recordings as reactions. As indicated by Kane et. al. [11], voice-banked phrases are often unused because they are hard to access in current AAC devices and are often too specific to be useful frequently. This novel application of voice-banking non-speech backchannel reactions would likely be usable across a wider range of scenarios than most banked phrases and would be simpler to access.

### The Voicesetting Editor

The key tradeoff in the design of the Expressive Keyboard is that it limits users' control over the exact properties of the output speech (i.e., emoji, punctuation, and ALM reactions perform a predefined set of operations) in order to ensure that adding expressivity to speech is fast and easy. While this tradeoff is important when users are engaged in synchronous, co-located communication, it becomes less important when users are preparing text to speak ahead of time (e.g., before a medical appointment, or before giving a public speech, or when preparing stored phrases for repeated use). In this asynchronous case, time is less of a factor, meaning that users may be willing to spend more time working on their speech if in return they have a higher degree of control over exactly how their text will be spoken by the AAC device. With this tradeoff in mind, we designed our second interface, the Voicesetting Editor (Figure 2a), to allow users to carefully craft the rendering of the synthetic voice. Clicking the "pen" icon key on the Expressive Keyboard opens the editor interface (Figure 1d), which allows the user to either edit the text they had just been composing in the keyboard or to load an existing text file from their device's file system.

The Voicesetting Editor parses input text into three types of tokens: words, punctuation, and vocal sound effects (derived from emoji in the input text) (Figures 2a and 2e). Tokens are displayed in reading order with enough padding to allow for sufficiently large gaze targets (minimum size of 60 x 60 pixels). Word and sound effect tokens are displayed as rectangular buttons while punctuation tokens are displayed as bubbles. Sound effect tokens display textual descriptions of the effect they represent (e.g., "Laugh" for laughter).

For each token, there is a set of voice properties that users can adjust through the editor interface. Several voice properties can be applied to all types of tokens: emotional tone, rate of speech, volume, and pitch. There are also several properties that can only be adjusted for particular token

types. Word tokens have an emphasis property to allow users to emphasize words, and punctuation tokens have a "pause" property to control the amount of silent time between the pronunciation of words. Importantly, SSML supports continuous ranges of values for many of these properties, but our editor only supports a fixed set of values for each property (Figure 2d). In early iterations, users were able to set continuous-valued properties to any value within the appropriate range, but this was found to be tedious (e.g., dwell-clicking a button 20 times to increase the volume from 0% to 100% in increments of 5%) and produced confusion over the meaning of different values (e.g., if there is a barely perceivable difference between a volume of 90% and 95%, users may feel that their change is not being respected by the system). Therefore, we chose a set of reasonable values for each property and assigned meaningful labels to each value. For instance, the volume property has the labels and values: "Very Quiet" (10%), "Quiet" (50%), "Loud" (100%), and "Normal" (a configurable default value).

The editor allows for setting properties both on individual tokens and over ranges of tokens. Dwell-clicking on a single token opens a radial menu surrounding that token (Figure 2.b). We use radial menus in the single-token editing case in order to reduce the distance the eye must travel to change properties of a token [8]. Dwell-clicking on the white circular button on the left side of the editor puts the editor into selection mode, allowing users to choose a range of tokens by dwell-clicking the first token in the range followed by the last token in the range. Making a range selection will open a horizontal linear menu in the center of the screen for editing properties over the entire range of tokens (Figure 2c).

Values of speech properties are represented visually in the editor interface in several different ways (Figure 2e-g). Font boldness of the text in word token buttons corresponds to the amount of emphasis placed on that word. Baseline height of the text in word token buttons corresponds to the pitch of the voice when speaking the word. Volume and speech rate are displayed as lines above the set of tokens sharing the same volume or rate property values. At the beginning of each of these lines, the property value is displayed in order to provide users with a direct method of inspecting the property. The only case in which these lines are not displayed is when the volume or rate property is set to its default value. Finally, the emotional tone of voice is displayed using a similar overline approach, encoding the tone of voice by displaying the emoji corresponding to that tone at the start of the line and coloring the line with a different color for each emotional tone.

While the Voicesetting Editor is necessarily more complex than the Expressive Keyboard, care was taken to minimize the time and effort required in the editor interface. A "play" menu option is available in every editing menu to ensure users can preview any changes they make before committing to the change. This allows for rapid exploration of the effects of potential edits. Additionally, a menu bar at the top of the editor (Figure 2a) allows for undoing and redoing changes, removing all changes, previewing the current synthesized speech for all of the text given all of the edits that have been made, and previewing the original synthesized speech for all of the text without any changes. Finally, in order to minimize the number of changes that need to be carried out in the editor, the Voicesetting Editor is connected to the Expressive Keyboard. Emoji and punctuation marks input through the Expressive Keyboard have their corresponding effects propagated into the Voicesetting Editor.

## EVALUATION

The evaluation of our voicesetting interfaces consisted of two parts. First, we evaluated the quality of the speech produced by each interface as compared to the output from a current AAC system, using an online questionnaire. Second, we collected usability information as well as qualitative feedback on both designs from people with ALS.

### Evaluation Prompts

In order to evaluate our voicesetting interfaces in the online questionnaire, we designed a set of prompts each consisting of text and markup describing how the text should be spoken for the purpose of recording actor-created audio clips as the ground truth for each prompt (Supplemental Table 1). To develop this set of prompts, we started with a set of 40 prompts and iteratively revised the set of prompts (via pilot testing the studies described below) until nine were selected as the final set. During this iterative process, we optimized for maximizing the variation between the expressive nature and tone of the prompts, while keeping the resulting lengths of the online questionnaire and lab study reasonable. In the final set of prompts, five were prompts drawn from a corpus of ambiguous English sentences (sentences that have multiple meanings, disambiguated by prosody when spoken) developed by Huenerfauth et. al. [10] (prompts 1-5), and four were developed manually in order to include prompts for which pacing of the speech is important (prompts 6, 8, and 9), and for which multiple emotions are expressed within a single sentence (prompt 7).

### Online Study: Speech Quality

#### *Methods*

We created an online questionnaire in order to evaluate the quality of the synthesized speech that can be created with the Expressive Keyboard and Voicesetting Editor as compared to the default synthesized speech produced by AAC devices. Each prompt described above was rendered to synthesized speech under three different conditions (using the prompt text without any of the script markup in brackets):

1. **Unedited Condition**: the default output from passing only the plain prompt to the synthesizer.
2. **Keyboard Condition**: the output from the Expressive Keyboard as created by an expert user (the first author).
3. **Editor Condition**: the output from the Voicesetting Editor as created by an expert user (the first author).

The questionnaire consisted of two sections. The first section was designed to evaluate whether listeners perceive the

| Prompt ID | Q1 | | | | Q2 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Fisher's Test (*p*) | Most Frequent Response | | | Median Response | | | Friedman's Test | | Kendall's | Wilcoxon Test (*p*) | | |
| | | U | K | E | U | K | E | $\chi^2$ | *p* | W | U - K | U - E | K - E |
| 1 | 0.0531 | Que. | Sad | Ang. | 2.5 | 2.5 | 2 | 3.287 | 0.1933 | 0.041 | *No Sig. Diff.* | | |
| 2 | ***4.97E-25*** | Neu. | Ang. | Ang. | 2 | 4 | 4 | 48.88 | ***2.43E-11*** | 0.611 | ***1.99E-07*** | ***4.37E-07*** | 0.8823 |
| 3 | ***3.98E-04*** | Neu. | Neu. | Neu. | 2 | 2.5 | 2 | 0.429 | 0.8071 | 0.005 | *No Sig. Diff.* | | |
| 4 | ***1.98E-11*** | Neu. | Ang. | Sar. | 2 | 2 | 4 | 30.24 | ***2.72E-07*** | 0.378 | 0.1126 | ***3.22E-06*** | ***4.79E-05*** |
| 5 | ***1.17E-11*** | Neu. | Ang. | Ang. | 3 | 3 | 3 | 2.941 | 0.2298 | 0.037 | *No Sig. Diff.* | | |
| 6 | 0.0432 | Sar. | Hum. | Sar. | 2 | 4 | 4 | 20.76 | ***3.11E-05*** | 0.259 | ***3.82E-04*** | ***1.13E-04*** | 0.5749 |
| 7 | *Not included in Q* | | | | 3 | 3 | 4 | 19.09 | ***7.17E-05*** | 0.239 | 0.2938 | **1.40E-03** | ***3.14E-04*** |
| 8 | 0.0205 | Que. | Neu. | Tho. | 3 | 3 | 5 | 32.05 | ***1.10E-07*** | 0.401 | 3.34E-02 | ***1.45E-06*** | ***2.10E-04*** |
| 9 | ***6.28E-06*** | Sad | Sad | Sad | 2 | 2 | 4 | 34.05 | ***4.04E-08*** | 0.426 | 0.2502 | ***1.23E-06*** | ***8.86E-06*** |

**Table 2. Online questionnaire results. Prompt IDs correspond to those in Supplemental Table 1. The abbreviations U, K, and E correspond to the Unedited, Keyboard, and Editor conditions, respectively. Q2 responses are on a scale from "Very Different" (1) to "Very Similar" (6). For all *p* value results, bold font indicates $p < 0.01$, while bold and underlined font indicates $p < 0.001$. Green cells indicate that the most frequent response matched the intended emotion. Post-hoc Wilcoxon tests (with Bonferonni correction, $p < 0.017$) were only performed in Q2 for prompts that had a significant *p*-value from Friedman's Test ($p < 0.01$).**

intended emotion and tone of the prompts better under any of the three conditions. The first section contained the following multiple-choice question (Q1) asked for each recording of each prompt with the exception of prompt 7 (8 x 3 = 24 questions): "*Which of the following words best describes the emotion and tone of the audio clip?*" with the possible responses "angry," "questioning," "thoughtful," "sad," "happy," "humorous," "sarcastic," "neutral," or "other." Prompt 7 was excluded from this section of the questionnaire so that all questions in this section had only a single intended ground truth emotion or tone.

The second section was designed to evaluate whether listeners perceive the speech output from any of the three conditions as being more accurate given a ground truth actor recording for each prompt. Actor recordings were used for the ground truth so that all participants were tasked with making the same comparison (e.g., some participants may conceptualize "angry" speech differently, but an actor recording of angry speech provides each participant with the same baseline). In a similar setup to the first section, the second section contained the following question (Q2) asked for each of the prompts (9 x 3 = 27 questions): "*How closely does the synthesized speech represent the emotion and tone of the real-world speech?*" with six-point Likert-type responses ranging from "Very Different" to "Very Similar". The presentation sequence of prompts and conditions within each questionnaire section were counter-balanced.

### *Participants*

The survey was distributed through a mailing list of employees at our organization, and 40 responses were collected. All participants were native English speakers, and none had hearing impairments. It is important to note that the use of non-AAC users for this study was appropriate given that this survey was evaluating listener perceptions of synthesized speech, and communication partners of AAC users would typically not themselves use AAC (e.g., family members, friends, professional caregivers, store clerks, etc.). In the next section ("Evaluation with People with ALS") we present results on the perceived output quality by AAC users.

### *Results: Speech Quality*

Table 2 shows an analysis of the results from the online questionnaire. Fisher's Exact Test was used to test for differences between the categorical distribution of responses to Q1 among the three conditions for each of the prompts. Significant differences were detected between the three conditions for each of the eight prompts ($p < 0.05$) with the exception of prompt 1 ($p = 0.053$). In the unedited condition, the most frequent Q1 participant response was the intended emotion in only one case (prompt 9). In the keyboard condition, the most frequent Q1 participant response matched the intended emotion for three prompts (prompts 2, 6, and 9). In the editor condition, the most frequent Q1 participant response matched the intended emotion for five prompts (prompts 1, 2, 4, 8, and 9).

With respect to Q2, results for each prompt fell into one of three groups. In the first group (prompts 1, 3, and 5) a Friedman test found no significant differences ($p > 0.05$) among conditions. In the second group (prompts 4, 7, 8, and 9) a Friedman test found significant differences among conditions and post-hoc Wilcoxon tests (with Bonferonni correction, $p < 0.017$) showed that there were no significant differences between the unedited and keyboard conditions, but synthetic speech from the editor condition was significantly more similar to the real speech than any other condition ($p < 0.017$ for each test). Finally, in the third group (prompts 2 and 6) a Friedman test found significant differences among conditions and post-hoc Wilcoxon tests show that there were no significant differences between the keyboard and editor conditions, but the synthesized speech from both the keyboard and editor conditions was significantly more similar to the real speech than the synthesized speech from the unedited condition ($p < 0.017$ for each test).

The questionnaire results indicate that both of our voicesetting systems create output with improved expressivity as compared to status quo AAC voice synthesis, and also that the Voicesetting Editor allows a larger and more nuanced range of expressivity than the Expressive Keyboard.

The findings also indicate there is much to be done in improving TTS engines (since some emotions had less-than-desirable clarity in their expression, and none yet achieve the quality of actor-recorded speech); however, TTS systems are not the focus of this research. Our system was built to leverage the features of current state-of-the-art TTS systems, and as the underlying TTS technology evolves in terms of realistic output for various SSML features, the quality of output achievable with our interface will also improve.

### Evaluation with People with ALS

Finally, in order to receive feedback from the target users of our voicesetting interfaces, we conducted two rounds of user testing with people with ALS (PALS). Because PALS often fatigue quickly, the first study was designed to be brief (about twenty minutes), focusing more on informal interface exploration and qualitative feedback; people who had the stamina and interest to complete longer study participated in a one hour session during which they performed more structured tasks. For the briefer, qualitative study, seven people with ALS participated (PL1 – PL7), four of whom are current users of gaze-based AAC devices; four people with ALS participated in the second, longer evaluation session (three of whom had done the shorter session prior; P1 – P4).

The goal of the first, briefer evaluation sessions was to collect qualitative feedback about the desirability of voicesetting interfaces overall and of the specific features included in our designs. Because of the time required to fully explore and understand the capabilities of the Voicesetting Editor, we focused this first evaluation session mainly on the Expressive Keyboard. After a tutorial, PALS could freely explore the Expressive Keyboard, including the emoji, punctuation, and ALM reactions; this free-form use allowed us to verify that the target user population could operate the interface successfully, and allowed PALS to experience the interfaces' capabilities so that they could provide qualitative feedback. The goal of the second, longer evaluation sessions was to enable participants to complete longer, more structured expressive editing tasks using both the Expressive Keyboard and the Voicesetting Editor, and to have participants explicitly compare and contrast the effort/quality tradeoff of these two designs.

#### *Evaluation with PALS Part 1: Qualitative Feedback*

Our first evaluation session with PALS consisted of three phases. In the first phase, participants were shown a video describing the Active Listening Mode feature of the Expressive Keyboard and then they tested out each of the expressive reactions in the ALM. After trying the ALM, participants were asked how often they would use the ALM as it is currently implemented, and how often they would use the ALM if the expressive reactions were voice-banked recordings of their own voice. Both questions were Likert-type items with responses: "Never," "Less Than Once per Week," "Multiple Times per Week," "Once per Day," "Multiple Times per Day," "Once or More per Hour." Participants were then asked how useful they found each of the currently available expressive reactions on a five-point Likert-type scale with responses from "Very Not Useful" to "Very Useful." To complete the first phase, participants were asked if there were any expressive reactions that were not included in the ALM that they would like to have included.

In the second phase, participants were shown a video describing the emoji operators of the Expressive Keyboard and then they tested the emoji on the phrase "Did you hear what James said? Only James would say something like that." (We used a two-sentence prompt so that participants could hear the contrast between the first sentence in the default voice, and the latter sentence to which they could apply emoji.) Participants had as much time as they desired to try out the interface, and typically explored for several minutes (i.e., using each of the emoji keys on either our sample sentence or their own sentences, and trying each of the ALM reactions). Participants were then asked how often then think they would use the emoji of the Expressive Keyboard to change the tone of their speech. Responses to this question were on the same scale as the frequency of use questions asked in the first phase of the study (ALM). Participants were then asked to rate the usefulness of each of the Expressive Keyboard's eight emoji on the same usefulness scale as used in the previous phase. We also asked participants if there were any emoji not included in the Expressive Keyboard that they would like to have included.

In the third phase, participants were shown a video describing the Voicesetting Editor and then they tested the editor on the same test prompt as in the previous phase Participants were asked how frequently they would use the editor to prepare speech on the same scale as in the previous two phases. In both the second and third phases, the prompt was automatically loaded into the interfaces, and participants were allowed to use the interfaces for as long as they wanted. Finally, participants were asked if they had any thoughts or feedback on any of the voicesetting interfaces.

#### *Results: PALS' Qualitative Feedback*

Table 3 summarizes participant responses about the usefulness of vocal sound effects in the ALM and emoji in the Expressive Keyboard. Participants indicated that laughter and "argh" were the most useful ALM reactions while the sarcastic scoff and the sharp intake of breath were rated least useful. Table 4 summarizes the frequency with which PALS anticipated using each of the interfaces. While participants were extremely enthusiastic about the concept of voicesetting in general, and particularly about the ALM, it is important to note that this sort of evaluation is known to be subject to positivity bias, and it is therefore more interesting to consider the relative differences between the responses for each interface (Table 4). While pleased by the level of control offered by the Voicesetting Editor, as expected, PALS reported they would likely use it less often than the keyboard, due to the time and effort involved, making it more appropriate for careful composition rather than real-time interactions.

| Effect | Median | Mean (SD) | Emoji | Median | Mean (SD) |
|---|---|---|---|---|---|
| Laugh | 5 | 4.6 (0.5) | Funny | 4 | 4.0 (1.4) |
| Argh | 4 | 4.0 (0.6) | Happy | 4 | 3.9 (1.3) |
| Ugh | 4 | 4.0 (0.8) | Sad | 4 | 3.9 (1.3) |
| Oh | 4 | 3.7 (1.4) | Irritated | 4 | 3.7 (1.4) |
| Hmm | 4 | 3.6 (1.4) | Sarcastic | 4 | 3.6 (1.5) |
| Cough | 3 | 3.4 (1.0) | Angry | 4 | 3.6 (1.5) |
| Pfft | 3 | 3.1 (1.1) | Surprised | 3 | 3.0 (1.5) |
| Breath | 3 | 3.1 (0.7) | Calm | 3 | 2.9 (1.3) |

**Table 3. PALS' ratings of usefulness of reactions in the Active Listening Mode and emoji in the Expressive Keyboard (sorted by mean response). Values are on a Likert-type scale from "Very Not Useful" (1) to "Very Useful" (5).**

In open-ended feedback, participants were interested in the potential to customize the reactions in the ALM in order to represent their personality. A family member of participant PL1 asked, "would his friends be able to record them?" referring to whether musician friends of PL1 could create custom reactions for him. PL1 indicated that he liked the idea and added that he would also want to include catchphrases in the ALM for comedic effect. PL2 echoed this sentiment, indicating he would want to include his catchphrase "Doood!" in the ALM. Several other reactions were also requested including "Uh-huh," "Doh!," "O-M-G," and "Eh?". PL4 indicated that he thinks the ALM could be useful because, "when there is a crowd, it is hard to interject" (due to the length of time needed to compose text via gaze-typing) and the ALM could serve as an interjection method.

With respect to the Expressive Keyboard, four of the seven participants commented on the subtleness of the differences between the different tones of the synthesized voice. PL7 said, "There's not a lot of difference between the different voices… the sounds [non-speech audio from the ALM and some of the emoji] have more impact than the different voices." PL2 also felt that the sound effects had more impact than the synthesized tone of voice: "What the emotion is is still hard to tell. The sound effects are helpful though." PALS also suggested several additional emoji, including "Shocked" (PL3 wanted "Shocked" to be more extreme than the current "Surprised"), and "Frightened" (PL7 wanted to use this when discussing possible health concerns).

For the Voicesetting Editor, PL1 indicated that it would be "great for speeches." PL1 described a charity event he was currently preparing to MC. In his capacity as MC, he had spent several weeks preparing a comedy routine with another PALS. He indicated that he would have liked to have the Voicesetting Editor during this preparation. PL4 said "saving the phrase would be helpful," describing a scenario in which he could edit and then save important phrases for reuse in the future. PL7's spouse indicated that they could imagine using the Voicesetting Editor in the infrequent case when he is "crafting a special communication."

### *Evaluation with PALS Part 2: Editor Comparison*

Our second evaluation session with PALS aimed to balance the difficulty of doing highly-controlled user studies with this group with the goal of getting more structured feedback on the tradeoffs between the effort required versus output quality produced for our two voicesetting interfaces. We employed a semi-structured approach in which each PALS performed similar tasks, but on their own unique content. While using personalized content for each participant makes cross-participant comparisons difficult, it adds more realism to the scenario; further, highly controlled studies and cross-participant comparisons are very difficult to run with this population due to factors such as fatigue during long sessions, the need to take frequent breaks to rest or adjust breathing or gaze equipment, and the difficulty of recruiting large enough populations for statistical significance

| Interface | Mode | Median | Mean |
|---|---|---|---|
| ALM + VB | 5 | 5 | 4.7 (0.8) |
| ALM | 5 | 5 | 4.6 (1.0) |
| Keyboard | 5 | 5 | 4.0 (1.7) |
| Editor | 3 | 3 | 3.0 (0.6) |

**Table 4. Frequency with which PALS expect to use each interface (sorted by mean response). Values are on a Likert-type scale from "Never" (1) to "More Than Once Per Hour" (6).**

After viewing a tutorial about the two systems, we asked PALS to compose a sentence they would typically say given a specific emotion (we repeated this exercise for each of three emotions: angry, happy, and sad). They first gaze-typed this sentence and played it in the default AAC rendering. Next, they could take as much time as needed to use the Expressive Keyboard to craft the target emotion; after playing that output, they answered Likert-type questions comparing and contrasting the default rendering to the Expressive Keyboard rendering. Then, they used the Voicesetting Editor on that same sentence, again trying to create the target emotional nuance. Afterwards, they again answered Likert-type questions contrasting this editor with the Expressive Keyboard and the default rendering. We also recorded the time PALS spent in each interface in order to create the desired effect (excluding time spent typing the prompt text). Note that we did not analyze text input errors as our evaluation is focused on the augmentation and markup of text for speech rather than the text input itself. Finally, PALS had the opportunity to provide open-ended feedback.

### *Results: PALS' Editor Comparison*

After completing each prompt (angry, happy, sad) in each condition (default, Expressive Keyboard, Voicesetting Editor), participants rated the similarity of the resulting TTS output to how they would say the phrase with their personal voice on a scale from 1 to 5 ("not similar at all" to "very similar"). In 10 out of 12 responses, participants rated the output from the Expressive Keyboard condition higher than the default output (1 response rated the output as the same, and 1 as less similar). In 7 out of 12 responses, participants rated the output from the Voicesetting Editor condition higher than the default output (3 responses rated the output as the same, while 2 responses rated it as less similar). When asked to choose if they would prefer the default output or the

output from the Expressive Keyboard condition, 10 out of 12 responses chose the Expressive Keyboard output. Similarly, when asked to make the same forced choice between the output from the Voicesetting Editor condition and the default output, 11 out of 12 responses chose the output from Voicesetting Editor. Finally, when participants were asked to choose which output they preferred most for each prompt, the output from the Voicesetting Editor was chosen 8 out of 12 times and the Expressive Keyboard 3 out of 12 times.

With regards to the time and effort required to use the Voicesetting Editor versus the Expressive Keyboard, the Voicesetting Editor did require more time and effort, as expected. On average, participants spent 2.81 (SD = 1.54) times longer in the Voicesetting Editor (mean = 182.2 sec) than in the Expressive Keyboard (mean = 73.6 sec). This timing data is inclusive of time spent exploring the available options and previewing all the feedback in each interface (but exclusive of text input time), so we expect the overall time to complete tasks during real world use to be significantly faster. When asked to rate the time and effort required to get the final output from each condition on a scale from "not at all worth the time and effort" (1) to "very worth the time and effort" (5), both the mode and median response for the Expressive Keyboard was 4, while the mode and median response for the Voicesetting Editor was 3.

All four participants conveyed strong feelings about the system in open feedback. P1 described how this system would provide him the ability to have his own voice again: "this IS very valuable work, there are few voice options today so this would be a way for me for the first time in five years to hear my own voice again!" On a similar note, P2 discussed the Voicesetting Editor saying, "[It] feels like you can add a bit more personality to it [the output speech]." Each participant described potential approaches for leveraging the power of the Voicesetting Editor while minimizing the required effort. All four participants noted that they wouldn't use the editor for regular speech but would use it for crafting and saving common phrases. P3 noted, "I would do more accurately in the real world," indicating that if he could save his edited phrases, he would spend more time crafting them to get them to sound just right. P1 proposed using the editor to set default states for the standard TTS output (e.g. setting the voice to have a particular emotion and rate by default). P2 wanted to be able to customize the effects of emoji by having the system automatically recognize common edits and adjust the effects of emoji based on that information.

## DISCUSSION

Through the design of our two voicesetting interfaces, we have explored approaches to allowing end-users of AAC devices to specify the expressive nature of the text they input to their system, and through this process, we have produced a better understanding of the limitations of the underlying TTS technologies that are currently limiting the expressivity that is possible through AAC. In our online questionnaire, the first set of prompts described in the results section highlight some of the limitations of state-of-the-art TTS engines. This set of prompts included an angry question dependent on heavy, properly-placed emphasis (prompt 1), a statement whose implication is completely dependent on a subtle use of emphasis (prompt 5), and a question that could sound like a statement if the speaker does not lift their pitch at the end of the sentence (prompt 3). For these three prompts, responses showed that listeners did not find any of the synthetic speech output conditions to be similar to the actor's recording (none of the three conditions had median responses greater than "Neutral" in terms of similarity to the actor-recorded speech). This indicates that TTS engines are not currently flexible enough with respect to control over emphasis or complicated prosodic processes like making a statement sound like a question, regardless of how accurate the input SSML is. This may indicate that the use of additional feedback modalities, such as the visual displays designed in the work of Sobel et. al. [25], are warranted for increasing the expressivity of AAC systems in situations where current expressive TTS techniques are insufficient; the interfaces and interaction styles that we introduced could also be used to drive an awareness display as an alternative or supplement to TTS output.

Another issue with the underlying TTS technology is that the emotional tones that TTS engines are currently capable of generating are not yet distinct or natural-sounding enough as described by the participants in our study with PALS. The spouse of PL6 noted that "intimate and frequent conversation partners like myself would learn to recognize the differences [between tones], but I suspect a more occasional conversation partner might not be able to discern them." The qualitative results from participants in our study with PALS are also supported by the responses we observed to Q1 in the online questionnaire. In the Expressive Keyboard and Voicesetting Editor conditions, the most frequent response was the intended emotion in only 50% of the questions (Table 2). In order to create AAC systems capable of producing the expressivity of natural human speech, TTS engines will necessarily need to improve the quality and range of tones they are capable of rendering.

These limitations of TTS technology notwithstanding, the current state of TTS technology still allows for the development of AAC systems that are capable of being far more expressive than the current state of the art. As described by PALS in our feedback sessions, and supported quantitatively by the online questionnaire results, the addition of vocal sound effects is a simple yet effective method of increasing the expressiveness of the synthesized speech (as either an alternative or supplement to modifying vocal tone via SSML). This is powerful because only a single dwell-click can provide an AAC device with enough information to render synthesized speech with appropriate sound effects, as demonstrated in the Expressive Keyboard. This result should be leveraged in the development of TTS engines as expanding the set of available sound effects would be a low-cost way to quickly improve expressivity.

In addition to the use of vocal sound effects in synthesized speech, PALS were excited about the use of sound effects in the Active Listening Mode of the Expressive Keyboard. PALS were enthusiastic about suggesting sounds they wanted to be able to play through the ALM, with suggestions reflecting the personalities of the PALS, including catchphrases and comedic interjections. The ease of playing these simple expressive reactions allows for users to express themselves in a far more natural and responsive manner than is currently allowed through AAC devices. Based on users' reactions to the ALM, we suggest encouraging PALS who have not yet lost speech production capabilities to voicebank non-speech audio (i.e., vocal sound effects) in addition to the standard process of voicebanking phrases.

### Extensions/Design Insights

The voicesetting interfaces we designed in this work certainly do not cover all aspects of human expressivity that are absent in current AAC. To that end, we believe the work presented here has the ability to act as a platform for further expressive AAC extensions. As an example, in the work of Kane et. al. [11], participants described the desire to be able to change voices, both wanting to be able to speak fluently in multiple languages (i.e., using multiple languages within a single utterance) and wanting to be able to change the persona of their voice (e.g., when reading their child a book, speaking differently for the different characters). Our design of the Voicesetting Editor currently focuses on allowing users to set the tone of the current voice; however, this same interface could be used to allow users to choose different voices altogether at both the phrase and word levels. Additionally, Kane et. al. [11] found that PALS wanted to be able to edit the pronunciations of words in their AAC device. While this functionality is currently available in some AAC devices, participants' comments reflected that this functionality is difficult to access and use. The Voicesetting Editor could be employed for this sort of word-level pronunciation editing. This extension would also address the comments that participants in our lab study expressed about wanting to be able to add vocal sound effects and prosodic modifications within words rather than between or across multiple words.

Outside of extending the current editing user experience, several other extensions to this work would be highly valuable. For instance, our use of emoji to signify the expressive speech properties of text was a simple and powerful markup technique that can easily be applied in a broad range of AAC systems; however, it is possible that the operation carried out by each emoji (currently static, based on heuristics) could be customized to the voice of each individual user by tracking and learning from the changes users make to the default operations carried out by emoji in the Voicesetting Editor. This machine learning approach to expressivity could be expanded to build models for automatically marking up expressivity in simple speech, reducing the input effort required by users.

### Limitations

Given the low incidence rate of ALS (1 in 50,0000) [15], and the extreme fatigue caused by the disease, recruiting a large number of participants for a study like ours is challenging. Further, issues of fatigue, communication difficulty, transportation, and medication timing (among others) make it difficult for people with ALS to participate in a lab study. While we received a significant amount of feedback from PALS throughout our iterative design process as well as in our summative studies, collection of in-depth usage information and design critique requires a longer-term deployment. For example, while PALS indicated they foresaw scenarios in which they would use the more complex Voicesetting Editor interface, long-term deployments with PALS would be necessary in order to determine whether PALS' ultimately find the effort/quality tradeoff of the Voicesetting Editor acceptable in situ. More generally, in order to fully evaluate the usability and utility of our voicesetting interfaces, long-term deployments with a greater number of PALS will be critical. Given that our system is currently limited by the capabilities of current TTS engines, evaluating or system using a Wizard-of-Oz approach in place of a TTS engine would allow for better understanding the impact voicesetting interfaces could have given improved TTS capabilities. Additionally, there are other user groups of gaze-based AAC devices that future evaluations could include (e.g., people with Spinal Muscular Atrophy, Muscular Dystrophy, Locked-In Syndrome, etc.). The design of AAC for children and teens can also benefit from increased expressivity, as noted by Light et al. [12]; exploring how to adapt and translate our interfaces for the unique needs and capabilities of youth with speech disabilities is an interesting area for further research.

## CONCLUSION

In this paper, we introduced the concept of *voicesetting interfaces* for allowing AAC users to explicitly control the expressive properties of synthetic speech. We contributed two novel designs for voicesetting interfaces that offer different effort/quality tradeoffs. The Expressive Keyboard treats emoji and punctuation as expressive operators over the surrounding speech, adding vocal sound effects, changing the tone of voice, and altering prosodic features such as the pitch of the voice. The Active Listening Mode of the Expressive Keyboard allows users to rapidly express reactions (i.e., vocal sound effects like laughter) while listening to others speak. The Voicesetting Editor provides users with the ability to carefully craft the way their text should be spoken. These designs were evaluated through an online questionnaire to evaluate the perceived quality of speech authored with these interfaces, and through in-lab qualitative and semi-structured feedback sessions with people with ALS.

## ACKNOWLEDGMENTS

We thank all of our participants for their time and feedback. We would also like to thank the members of Microsoft Research's Enable team for their assistance with this work.

## REFERENCES


1. Matthew P. Aylett, Graham Pullin, David A. Braude, Blaise Potard, Shannon Hennig, and Marilia Antunes Ferreira. 2016. Don't Say Yes, Say Yes: Interacting with Synthetic Speech Using Tonetable. In *Proceedings of* CHI EA '16. ACM, 3643-3646.
2. Daniel C. Burnett, Mark R. Walker, and Andrew Hunt. 2004. Speech Synthesis Markup Language (SSML) Version 1.0. W3C. Retrieved September 9, 2016 from http://www.w3.org/TR/speech-synthesis/
3. Nick Campbell. 2007. Approaches to conversational speech rhythm: Speech activity in two-person telephone dialogues. In *Proceedings of the 16th International Congress of the Phonetic Sciences*, Saarbrucken, Germany, pp. 343-348.
4. CereVoice Engine Text-to-Speech SDK. 2016. CereProc. Retrieved September 9, 2016 from https://www.cereproc.com/en/products/sdk
5. Marcela Charfuelan and Ingmar Steiner. 2013. Expressive Speech Synthesis in MARY TTS Using Audiobook Data and EmotionML. In *Proceedings of Interspeech* 2013. ISCA.
6. Paul Ekman, E. Richard Sorenson, and Wallace V. Friesen. 1969. Pan-cultural elements in facial displays of emotion. In Science 164.3875: 86-88.
7. Susan Goldin-Meadow. 1999. The role of gesture in communication and thinking. In *Trends in Cognitive Sciences*, 3(11), 419-429.
8. John P. Hansen, Anders S. Johansen, Michael Donegan, David J.C. MacKay, Phil Cowans, Michael Kühn, Richard Bates, Päivi Majaranta, and Kari-Jouko Räihä, K. J. (2005). D4. 1 Design Specifications and guidelines for COGAIN eye-typing systems. Retrieved September 9, 2016 from http://wiki.cogain.org/images/a/a7/COGAIN-D4.1.pdf
9. D. Jeffery Higginbotham. 2010. Humanizing Vox Artificialis: The Role of Speech Synthesis in Augmentative and Alternative Communication. In *Computer Synthesized Speech Technologies: Tools for Aiding Impairment,* 50.
10. Matt Huenerfauth, Pengfei Lu, and Andrew Rosenberg. 2011. Evaluating Importance of Facial Expression in American Sign Language and Pidgin Signed English Animations. In *Proceedings of* ASSETS '11, 99-106. http://dx.doi.org/10.1145/2049536.2049556
11. Shaun K. Kane, Meredith Ringel Morris, Ann Paradiso, and Jon Campbell. 2017. “At times avuncular and cantankerous, with the reflexes of a mongoose”: Understanding Self-Expression through Augmentative and Alternative Communication Devices. In *Proceedings of* CSCW ‘17.
12. Light, Janice, Rebecca Page, Jennifer Curran, and Laura Pitkin. 2007. Children's ideas for the design of AAC assistive technologies for young children with complex communication needs. In *Augmentative and Alternative Communication.* 23, 4: 274-287.
13. Päivi Majaranta and Kari-Jouko Räihä. 2002. Twenty years of eye typing. In *Proceedings of* ETRA ’02, ACM Press, 15. http://doi.org/10.1145/507072.507076
14. Mary Text-To-Speech. Retrieved September 13, 2016 from http://mary.dfki.de/
15. Hiroshi Mitsumoto. 2009. Amyotrophic lateral sclerosis: a guide for patients and families. Demos Medical Publishing.
16. Martez E. Mott, Shane Williams, Jacob O. Wobbrock, and Meredith Ringel Morris. 2017. Improving Dwell-Based Gaze Typing with Dynamic, Cascading Dwell Times. In *Proceedings of CHI ’17*.
17. Esther Nathanson. 2016. Native voice, self-concept and the moral case for personalized voice technology. In *Disability and Rehabilitation*. 1-9.
18. Nuance Loquendo. 2016. Nuance. Retrieved September 9, 2016 from http://www.nuance.com/for-business/by-solution/customer-service-solutions/solutions-services/inbound-solutions/loquendo-small-business-bundle/tts-demo/english/index.htm
19. Sandra Pauletto, Bruce Balentine, Chris Pidcock, Kevin Jones, Leonardo Bottaci, Maria Aretoulaki, Jez Wells, Darren P. Mundy, and James Balentine. 2013. Exploring Expressivity and Emotion with Artificial Voice and Speech Technologies." In *Logopedics Phoniatrics Vocology* 38, 3: 115-125.
20. John F. Pitrelli, Raimo Bakis, Ellen M. Eide, Raul Fernandez, Wael Hamza, and Michael A. Picheny. 2006. The IBM expressive text-to-speech synthesis system for American English. In *IEEE Transactions on Audio, Speech, and Language Processing* 14, no. 4: 1099-1108.
21. Colin Portnuff. 2006. Augmentative and Alternative Communication: A Users Perspective. Lecture delivered at the OHSU, August 18, 2006. http://aac-rerc.psu.edu/index-8121.php.html
22. Graham Pullin and Shannon Hennig. 2015. 17 Ways to Say Yes: Toward Nuanced Tone of Voice in AAC and Speech Technology. In *Augmentative and Alternative Communication*, 31(2), 170-180.
23. Klaus R. Scherer and Heiner Ellgring, 2007. Multimodal Expression of Emotion: Affect Programs or Componential Appraisal Patterns? In *Emotion*, vol. 7, pp. 158-71.
24. Marc Schröder. 2009. Expressive speech synthesis: Past, present, and possible futures. In *Affective Information Processing*, pp. 111-126. Springer.

25. Kiley Sobel, Alexander Fiannaca, Jon Campbell, Harish Kulkarni, Ann Paradiso, Ed Cutrell, Meredith Ringel Morris. 2017. Exploring the Design Space of AAC Awareness Displays. In *Proceedings of CHI '17*.
26. System Usability Scale. Retrieved September 11, 2016 from https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html
27. Éva Székely, Zeeshan Ahmed, João P. Cabral, and Julie Carson-Berndsen. 2012. WinkTalk: a demonstration of a multimodal speech synthesis platform linking facial expressions to expressive synthetic voices. In *Proceedings of* SLPAT '12. 5-8.
28. Watson Text-To-Speech. 2016 IBM. Retrieved September 9, 2016 from http://www.ibm.com/watson/developercloud/text-to-speech.html
29. Sheida White. 1989. Backchannels across cultures: A study of Americans and Japanese. *Language in society*. 18, 01: 59-76.
30. Xiaoyi Zhang, Harish Kulkarni, and Meredith Ringel Morris. 2017. Smartphone-Based Gaze Gesture Communication for People with Motor Disabilities. To appear in *Proceedings of CHI '17*.