Tecla Sound: Combining Single Switch and Speech Access

Hemanshu Bhargav[1(✉)], Ijaz Ahmed[2(✉)], Margot Whitfield[1(✉)], Raaga Shankar[1(✉)], Mauricio Meza[2(✉)], and Deborah I. Fels[1(✉)]

[1] Ryerson University, 350 Victoria St., Toronto, Canada
{hbhargav, margot.whitfield, raaga.shankar, dfels}@ryerson.ca
[2] Komodo OpenLab, Toronto, Canada
ijaz.ahmad@ryerson.ca, mauricio@kmo.do

Abstract. Using single switch scanning with a mobile Smartphone can often be frustrating and slow to use. Phone functions, such as answering a call, either have specific timeout imitations that are set by service providers or may not be accessible using non-standard access methods. Single switch scanning can take longer than these timeout settings allow, resulting in missed calls or becoming stuck in endless automated call option menu loops. The Tecla single switch-scanning system has been augmented with a simple speaker dependent, limited vocabulary, offline speech recognizer to offer users one solution to this dilemma. Users decide on and record applicative words for specific phone functions, such as "answer," that can then be used instead of selecting that item from a scanning array. Given that speech production is often faster than selection from a scanning array, the user will be able to avoid the timeout limitation of their smartphone.

Keywords: Single switch use · Switch access · Accessibility speech recognition · Mobile devices

1 Introduction

Single switch interfaces are used by individuals who can only produce a single intentional and consistent physical action to activate a control [1]. Individuals with quadriplegia and other motor impairments, Cerebral Palsy, Multiple Sclerosis, and other diagnoses, require alternative access to computers and mobile devices [2]. Speech impairments may also be present in individuals with these diagnoses [2, 3]. The single switch interface involves one switch that is coupled with a method of presenting possible options, one at a time, often as a scanning array. The user presses the switch to indicate a selection once the desired option appears. The types of applications that are commonly used with single switches and scanning arrays are: environmental and wheelchair controls; general computer hardware and software; and specialized or dedicated alternative access systems such as speech generation devices. However, dedicated or specialized access systems can be expensive and have limited functionality. Instead, being able to access mainstream applications allows users to learn and

K. Miesenberger et al. (Eds.): ICCHP 2020, LNCS 12377, pp. 348–354, 2020.
https://doi.org/10.1007/978-3-030-58805-2_41

gain experience with those applications rather than adopting a specialized application which can be more expensive, inflexible or less readily available.

Mainstream desktop (e.g., macOS) and mobile operating systems (e.g., iOS and Android) now incorporate switch access accessibility features [4], which enable access to individuals who are unable to use standard interfaces like touch screens, keyboard or pointer devices.

While mobile devices are now, more than ever, natively accessible to single switch/scanning devices. However, using a single switch remains a very slow process. Todman [3] reports that typical word entry rates for single switch users are 2–5 words per minute, even with word prediction (compared to typical touch-typing keyboard rates of 20–45 words per minute). Assistive voice recognition with word prediction can reduce the number of required keystrokes by as much as 69% [5, 6].

If additional methods can be found to augment single switch technologies, which do not involve more physical movement, then the time it takes to access functionality, options or enter selections may increase. One relatively new and recently more robust technology is automatic speech recognition (ASR). As Rudzciz [7] states, "Since dysarthric speech is often 10 to 17 times slower than normal, ASR is seen as a viable alternative to improve communicatively in computer-assisted interaction" [pp. 248].

However, built-in ASR systems (voice assistants) on smartphones are designed for clear and unambiguous speech production and require an Internet connection. They also rely on remote data sets to perform recognition and do not allow users to personalize the voice models (such as what is used with speaker dependent systems), which can improve recognition [8]. A single switch user may be disconnected from the Internet or have mobile data deactivated, which then disables their smartphone's onboard speech recognition system. However, speaker dependent systems, such as Dragon Naturally Speaking, can be expensive and are often not intended to be integrated with single switch interfaces. Thus, any ASR system that complements single switch use must be accurate, inexpensive and robust without Internet connectivity.

In this paper, a Bluetooth enabled, limited vocabulary ASR application for use with the Tecla single switch interface is described. A use-case example is also presented to describe how the system could be used by single-switch users.

2 System Description

2.1 Switch Scanning and Mobile Devices

In mobile devices, switch accessibility features provide access to native, standard functionality. When a single switch scanning user receives a phone call, the scanning frame scans through all elements of call answering functions, one at a time. This is a time sensitive task and depending on the scanning settings of the user, the call may be missed before the answer button is scanned by the system. While some voice assistants in mobile devices may allow the user to answer a phone call, there is no option to hang up by voice, as the microphone is in use by the phone app during a call. This could lead to users to become "stuck" in corporate menus or voicemail systems, if they do not have alternative access to their device such as switch access.

2.2 Speech Interface Augmenting Tecla's Single Switch Functions

As a result of difficulties with the time restrictions for answering and hanging up functions experienced by Tecla users, we wanted to augment the single-switch system with a limited functionality speech interface. However, an important criterion was to implement a limited-phrase offline speech recognizer, so that Tecla users do not require Internet access in order to use it. In addition, this system could be used in parallel to phone use. Limited word/phrase encoding that can be user-specified, called a "hot-word," is possible as there are only a few time-dependent phone functions such as "answer" or "hang-up the phone".

There were a number of possible options for offline natural language processers (NLPs) including Picovoice [9] PocketSphinx [10], Snowboy [11], Kaldi [12] and Mozilla's DeepSpeech [13]. We selected the Snowboy system as it uses a neural network, which classified true/false for "wake-word" detection, allows developer modifications, and was optimized for limited hotword detection. Wake-word detection was necessary for mode-switching, so that hotwords were not confused with speech utterances that could occur in another application such as a phone conversation. Other offline NLPs required enterprise licensing (Picovoice), had disproportionately high false positives [14, pp.67] and poor "word-error rate" scores [10], did not provide wake-word support (Kaldi and Mozilla DeepSpeech), or were not intended for consumer deployment (Kaldi).

As shown in Fig. 1, the Tecla speech interface begins with an "initial setup" which prompts the user to configure a set of predefined phrases. The user begins by specifying their language choice, age, speech pattern, and microphone settings, as these are required by Snowboy in order to optimize the speech detection algorithm for particular technical voice parameters. The user is then presented with 15 default phrases such as answer, which can be used, edited or replaced (see Fig. 2).

To qualify as a hotword, restrictions are placed on user input such that commands are a minimum of ten characters in length, so that voice commands are sufficiently distinct in sound. To train the system, the user is prompted to record three samples voicing the phrase using whichever words/sounds that they can repeat. Once successful training occurs, a Snowboy "personal model" is generated for each voice command, specific to each user (see Fig. 1). This process is repeated until the user has trained the system for all of the desired hotwords.

To reduce the effects of variability in dysarthric speech, users are encouraged to record voice samples for their commands at various times of the day, each time assigning a command which is similar in enunciation to previous commands, which serve the same function. As the Snowboy system only permits three samples for training, this process will ensure that the "personal model" is not corrupted from environment noise and so that variations of the command that a user with dysarthric speech may speak are included in the model.

Audio samples must be recorded at frequency of 16 kHz to ensure the Snowboy recognizer accepts them for training [11]. Although most USB microphones can record at this frequency, mass-market, open source microphones, which comply with the Tecla Shield 3.0's small space, are fewer. As such, the Seeed Sound Dac for speech recognition is the default microphone for the system and functions independently from

the phone microphone. Gain and sensitivity for the microphone are user adjustable and can be adjusted to accommodate different levels of speech production quality—including dysarthric speech.

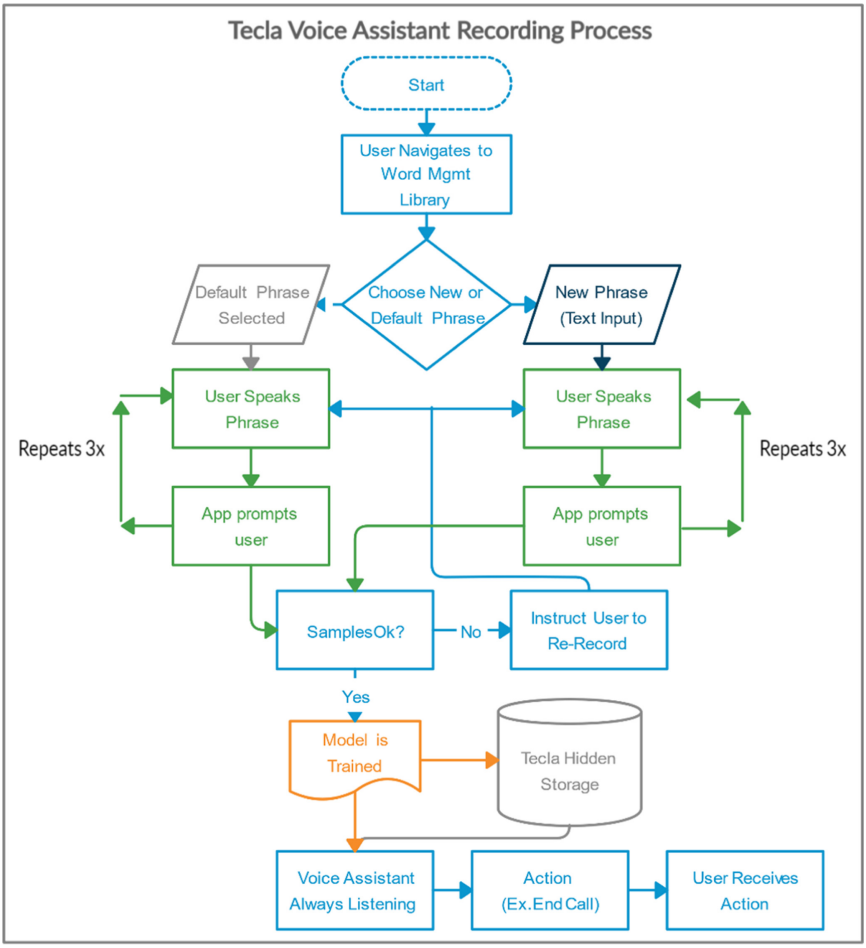

Fig. 1. Data flow during recording.

3 Use Case/Example Scenario

Instead of using the Tecla single switch/scanning solution for all phone actions, using voice commands would allow a user to replace some physical hardware/relay switch functionality with voice commands. The following scenario of the receipt and answering of a phone call on a user's smartphone is outlined and compared between the "single-switch only" version of Tecla and the speech augmented version.

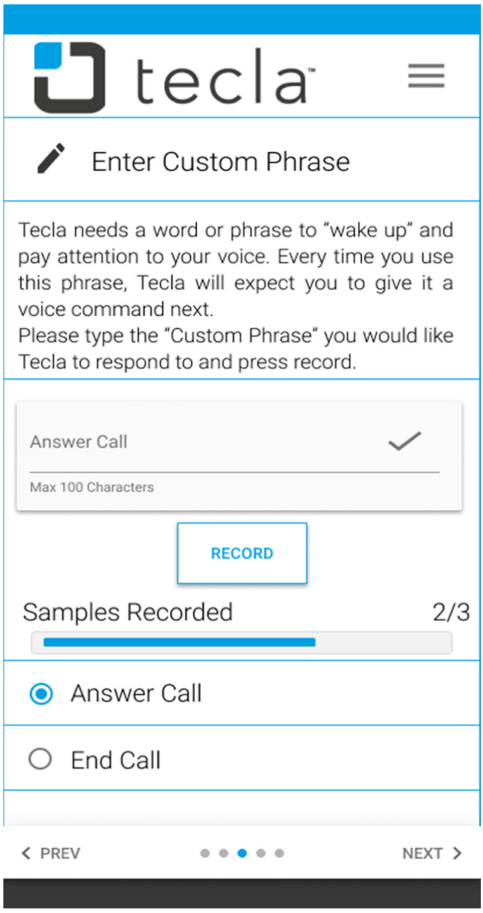

Fig. 2. Tecla Recording App Interface. Material Icons by Google, used under Apache License Version 2.0.

To use the single-switch scanning Tecla system, a user must press a switch to start the scanning functionality, wait until the desired interface element is highlighted by the system, and then press the switch again. Most smartphones are configured by the cellular service provider to ring for a specific number of times (e.g., for the Rogers service provider, the number of rings is nine) [15]. Once that time is exceeded, the phone reverts to voicemail. The time to activate the "answer call" item on the scanning menu is longer than allowed by the service provider. It is thus difficult to answer the phone within the designated number of rings, and calls are often missed. To perform the same tasks with the speech recognition system, the user would voice the command used for the "answer" function. This takes only the time to voice that single command.

While the number of commands that have been selected for use with the Snowboy offline recognizer is 15, it is sufficient to accommodate a number of phone-specific functions that are difficult to complete with single-switch scanning. In addition to

answering or ending a call, a command could be set to perform a gesture such as turning a page, or scrolling, or used as a modifier to use a switch with a secondary function. For example, the new Xbox adaptive controller uses switches as inputs to play games and allow users to create different layouts. Similarly, a voice command on the Tecla system could be used to allow users to switch between layouts – a switch between racing and adventure games is one possible scenario.

3.1 Limitations

The current speech recognizer is an open source tool for personal use, which is subjected to market interest and support from the community of developers. If the development community no longer wishes to support the tool, then there will be a potential risk to future development, troubleshooting or support. A different speech recognizer must be considered if this occurs.

4 Future Work

The next steps in this research is to carry out user studies with current Tecla single-switch-scanning users to determine the acceptability, usability and usefulness of the Tecla scanning/speech system for managing telephone functions in a smartphone. Time to access the various phone functions will be a key variable. The target user group must include users with varying levels of speech production ability, articulacy, severe dysarthria and environments with varying noise levels.

Acknowledgements. Funding is generously provided by the Accessible Technology Program of Innovation, Science and Economic Development Canada.

References

1. Koester, H.H., Simpson, R.C.: Method for enhancing text entry rate with single-switch scanning. J. Rehabil. Res. Dev. **51**(6), 995–1012 (2014)
2. Koester, H.H., Arthanat, S.: Text entry rate of access interfaces used by people with physical disabilities: a systematic review. Assistive Technol. **30**, 151–163 (2018). https://doi.org/10.1080/10400435.2017.1291544
3. Todman, J.: Rate and quality of conversations using a text-storage AAC system: single-case training study. Augmentative Altern. Commun. **16**, 164–179 (2000). https://doi.org/10.1080/07434610012331279024
4. Apple Inc.: Use Switch Control to navigate your iPhone, iPad, or iPod touch. https://support.apple.com/en-ca/HT201370
5. Swiffin, A., Arnott, J., Pickering, J.A., Newell, A.: Adaptive and predictive techniques in a communication prosthesis. Augmentative Altern. Commun. **3**, 181–191 (1987). https://doi.org/10.1080/07434618712331274499
6. Matiasek, J., Baroni, M., Trost, H.: FASTY—a multi-lingual approach to text prediction. In: Miesenberger, K., Klaus, J., Zagler, W. (eds.) ICCHP 2002. LNCS, vol. 2398, pp. 243–250. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45491-8_51

7. Rudzicz, F.: Production knowledge in the recognition of dysarthric speech (2011). http://search.proquest.com/docview/920144730/abstract/B15CC1C7A3F437CPQ/1
8. Nuance Communications: Dragon Speech Recognition - Get More Done by Voice. https://www.nuance.com/dragon.html
9. Embedded Wake Word & Voice Commands - Picovoice. https://picovoice.ai/products/porcupine/
10. Huggins-Daines, D., Kumar, M., Chan, A., Black, A.W., Ravishankar, M., Rudnicky, A.I.: Pocketsphinx: a free, real-time continuous speech recognition system for hand-held devices. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, vol. 1 (2006). https://doi.org/10.1109/ICASSP.2006.1659988
11. Kitt, A.I., Chen, G.: Snowboy, a Customizable Hotword Detection Engine—Snowboy 1.0.0 documentation. http://docs.kitt.ai/snowboy/
12. Povey, D., et al.: The Kaldi speech recognition toolkit
13. Mozilla: Welcome to DeepSpeech's documentation!—DeepSpeech 0.7.3 documentation. https://deepspeech.readthedocs.io/en/v0.7.3/?badge=latest
14. Rosenberg, D., Boehm, B., Stephens, M., Suscheck, C., Dhalipathi, S.R., Wang, B.: Parallel Agile – Faster Delivery, Fewer Defects, Lower Cost. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-30701-1
15. Rogers Communications: Set up additional features for your Home Phone voicemail - Rogers. https://www.rogers.com/customer/support/article/set-up-additional-features