THE CARES CORPUS: A DATABASE OF OLDER ADULT VOICES FOR DEVELOPING SPEECH RECOGNITION SYSTEMS

Victoria Young[1,2] and Alex Mihalidis[1,2,3]

[1] Graduate Department of Rehabilitation Science, University of Toronto;
[2] Institute of Biomaterials & Biomedical Engineering, University of Toronto;
[3]Toronto Rehabilitation Institute, Toronto.

ABSTRACT

To improve the automatic speech recognition component used within an intelligent, speech-based personal emergency response system, a collection of training and test speech was required of primarily older adults in emergency situations. This paper describes the development of the Canadian adult regular and emergency speech (CARES) corpus. The CARES corpus consists of a collection of read and spontaneous speech recordings from mainly adult actors aged 23-91 years with an emphasis on emergency type dialogue. A total of 40 participant voices are included in the corpus. More than 70% of the voices are adults over 50 years of age.

INTRODUCTION

Background

Personal emergency response systems (PERS) are commonly installed in the homes of older adults to provide access to immediate 24 hour emergency assistance when needed at the push of a button; for example, in the case of a fall or medical complication. The button is typically worn on the body, around the neck or wrist. PERS have been shown to support aging-in-place, lower caregiver and user anxiety, and decrease overall healthcare costs [3, 4]. However, despite high user satisfaction, a large proportion of PERS owners do not use their systems when needed. Reasons for non-use are varied but include sensitivity or burden from having to wear or remember to wear the button; potential loss of independence from outcomes related to falling or hospitalization; inability to press the button either because they do not want to or they cannot push it [1, 4, 5].

In addition, the largest proportion of calls to personal emergency response call centres actually consist of false alarms (accidental button presses) which may lead to unexpected calls to the subscriber, loss of work hours for family responders, and increasing workload for already stressed emergency care providers [4, 5]. In order for potentially life saving assistive technologies such as the PERS to be successfully adopted by older adults for aging-in-place, it is important that these systems be made accessible, usable, efficient and effective.

Research Focus

Our research focuses on investigating hands-free methods of initiating emergency response through the use of an intelligent, speech-activated PERS. See [2] for further details. This new technique would eliminate the need for body worn activators and could also address some of the technology use issues surrounding the traditional push-button PERS.

To improve the robustness and accuracy of the automatic speech recognition (ASR) component used within the intelligent, speech-based PERS, a collection of speech samples was required for system training and testing from the target population, preferably in various emergency situations. See [7] for further background on ASR and older adults.

Since a database of adult voices, in particular older adult voices, in Canadian English within an emergency context was not readily available, we decided to develop the **C**anadian **A**dult **R**egular and **E**mergency **S**peech (CARES) corpus. The CARES corpus includes a collection of read and spontaneous speech, as well as emergency words, phrases, and enacted dialogue collected from adults, the majority of which were over 50 years of age. Since true emergency situations could not be feasibly re-

enacted, voice samples from adult actors were targeted in mock emergency situations. This paper describes the development process of the CARES corpus for the purpose of testing and training an intelligent, speech-based PERS.

METHODOLOGY

The methodology used in this study has been reviewed and approved by the University of Toronto Ethics Board.

Participant Recruitment

Adult actors were targeted for recruitment within three age groups from the greater Toronto area (GTA). The three age groups consisted of: (1) adults 19+ to <55 years of age; (2) adults 55 to 69 years of age; and (3) adults 70 years of age and over. All actors were required to have a minimum of one year of prior acting experience with an acting group.

Actor participants were recruited from local theatre events (e.g., Toronto Fringe Festival), a senior's acting group (Act II) based out of a local university (Ryerson University), and from a "Performing Arts Lodge" located within the GTA which is a residence for individuals in the performing arts. Participants were also recruited via word of mouth from other participants. A minimum of five participants of each gender (male and female) were recruited for each age group category.

Participant suitability was determined through a telephone interview conducted prior to acceptance into the study. Participants were required to be Canadian residents with no or minimal accent; no motor speech difficulties; no or corrected hearing loss; no or corrected vision loss, medically stable, mobile and cognitively capable of consenting to participate in the study.

Recording Procedure

The entire speech recording session was designed to last approximately two hours. Participants were required to perform four different speaking exercises:

1. **Five minutes of free speech:** participants were asked to speak naturally and spontaneously about a miscellaneous topic of their choosing. If they had difficulty, questions were asked to facilitate the dialogue.

2. **Sentence Reading:** participants read a collection of phonetically rich and compact sentences (50 and 46 sentences respectively) derived from the database SCRIBE [6]. All sentences were pre-selected by the researchers and presented on a computer monitor.

3. **Emergency phrase and word speaking:** participants were instructed to read and then speak with emotion, short pre-compiled emergency phrases (185 phrases). A pre-selected word from this phrase was then repeated five times in different manners of speaking: normally, loudly, softly, quickly and slowly. All phrases and words were displayed on a computer monitor and participants were provided prompts for the five different manners of speaking. For the "slow" manner of speaking, participants were instructed to imagine they had difficulty forming their words.

4. **Enacting three emergency scenarios:** participants were assigned three of nine possible short emergency scenarios, all of which involved dialogue that might occur after pressing an assist button on a PERS. The three scenarios included: one accidental button push for assistance, one fall incident and one request for medical assistance. All scenarios were pre-written and pre-assigned by the researchers. Scenarios were provided to the participants for review and practice prior to the day of their voice recording.

The emergency phrases, words and scenarios were derived from actual, live, personal emergency response calls obtained from a local, private personal emergency response company (name not provided for reasons of confidentiality).

For speaking exercise 2, 200 phonetically rich sentences and 460 phonetically compact sentences from SCRIBE [6] were used. The phonetically rich sentences were divided into four sets of 50 sentences. The phonetically compact sentences were divided into ten sets of 46 sentences. Each participant was randomly assigned one set of sentences from both the phonetically rich and compact sentence groups.

For speaking exercise 3, the lists of emergency words and associated phrases were

first randomized and arranged into 'set 1'. This list was then placed in reverse order into 'set 2'. Each participant was assigned either set 1 or set 2 in an alternating pattern (one participant viewed the words in alphabetical order).

For speaking exercise 4, a total of three accidental push-button scenarios (A), three fall incident scenarios (F), and three medical assistance scenarios (M) were devised. A total of 27 scenario combinations resulted and these were placed in random order. Each participant was assigned one of the randomized 27 scenario combinations. This order was repeated when all 27 scenario combinations had been performed. The 3 scenario type combinations (e.g., A, F and M) were also arranged into 6 order combinations (e.g., A-F-M; or F-M-A; or M-A-F, etc.). Each participant was assigned one of the scenario type combinations. This order was repeated when all 6 scenario type combinations were completed.

Recording Environment

All speech recordings were conducted at the University of Toronto in quiet background conditions inside a double doored, sound attenuated booth of approximately 74 x 74 x 78.5 inches (DxWxH) in size.

Participants were seated inside the sound attenuated booth in front of a computer monitor. Headphones were worn over their ears in order to communicate with the researcher. A free standing microphone was used for the speech recordings positioned a few inches in front and to the left of the participant's mouth.

The researcher was positioned in a separate sound attenuated booth. Communication was via microphone and headphones with visual communication through a window adjoining the two sound booths.

Recording Equipment

Speech recordings were made using ProTools TDM Software on a dedicated Apple Computer (MAC OS X version 10.4.11, 3 GHz Dual-Core Intel Xeon). The microphone pre-amp was a Digidesign "PRE" and the audio interface was the Digidesign "192 I/O".

Participant speech was recorded at a sampling rate of 96 kHz and 24 bits.

The participant used an AKG Acoustics k271 studio headphone and an Audio-technica 4050 multi-pattern condenser microphone.

RESULTS

Participant Recruitment

A total of 40 participants, 19 male and 21 female, were recruited for the study over a six month period. Thirteen of the participants fell within the 19+ to <55 years of age group, twelve participants were in the 55 to 69 years of age group, and fifteen participants were in the 70 years of age and older group. See Table 1 for a breakdown of the participants by Age Group and Gender.

Table 1: Participants by Age Group and Gender

Age Group (years)	Gender	
	Male	***Female***
19+ to <55	6^	7
55 to 69	6	6*
70 and over	7&	8~
Gender Totals	19	21

^1 non-actor; &1 with minor accent; *1 with minor accent, 1 non-actor; ~3 with minor accent, 2 non-actors.

Nineteen of the participants had 15 years or more experience in the acting profession. Four participants included in the study had no acting experience and five participants spoke with a minor accent. Minor accents included British English, French and German. Participants spanned an age range from 23 to 91 years of age. See Table 2 for a breakdown of the participants by age range.

Fifteen of the participants were born outside of Canada. They represented the following countries: Austria, England, France, Germany, Italy, Japan, Scotland and USA. For participants born within Canada, the Canadian birth provinces included Alberta, British Columbia, Ontario, Newfoundland, and Nova Scotia.

Speech Recording Summary

Each participant completed all four speech exercises described in the Methodology Section, except for one participant who did not complete the emergency scenario exercise. A total of ~3,200 minutes of speech was recorded.

Table 2: Participants by Age Range

Age Range (years)	Gender	
	Male	*Female*
20's	2	3
30's	3	3
40's	1	0
50's	3	2
60's	3	5
70's	5	6
80's	2	1
90's	0	1

DISCUSSION

The length of time required for speech recording was approximately two hours; however, timing was dependent on how quickly the participants spoke and how many breaks were required. Younger participants tended to finish the study in less than two hours, while some of the older participants required more time to finish the exercises.

In the 3rd speaking exercise, for the word repetition component, the "slow" manner of speaking was interpreted in different ways. Variations included lengthening the word slightly, exaggerating the length, stuttering, and slurring the speech.

Due to their age and life experiences, the older participants were observed to be more realistic at portraying older adults in emergency situations especially for certain conditions (e.g., stroke, weakness, heart attack).

The CARES corpus is of similar size to the "few talker" set in the SCRIBE [8]. Although the speech sample size may not be large enough to train a large vocabulary ASR, the number of speech samples should be sufficient for training and testing the intelligent PERS ASR and preliminary field testing. If required, additional speech samples can be added in the future.

CONCLUSIONS

A collection of Canadian adult regular and emergency speech has been developed containing speech samples from 40 adults. Participants included mainly adult actors who were required to carry out four speech exercises including spontaneous and read speech, and enacted emergency dialogue, phrases and words. The CARES corpus contains roughly 3,200 minutes of speech. This corpus was primarily designed to further develop the ASR component of the intelligent PERS for older adults. It may also find uses in future research studies involving natural speech processing and computational linguistics.

ACKNOWLEDGEMENTS

The authors acknowledge the many people who helped support this project: the study participants; the Department of Speech Language Pathology and the Toronto Rehabilitation Institute's (TRI) Communication Team for lending the sound attenuation booth and recording equipment; the Intelligent Assistive Technology and Systems Lab; and the PERS company for providing the live emergency calls used in the study design. Funding has been provided by: the University of Toronto; the CIHR Health Care, Technology, and Place Fellowship (FRN:STP53911); Engineers Canada & TD Meloche Monnex; NSERC; and TRI.

REFERENCES

[1] B. Heinbüchner, M. Hautzinger, C. Becker, and K. Pfeiffer, "Satisfaction and use of personal emergency response systems," *Zeitschrift fur Gerontologie und Geriatrie,* vol.43, pp.219-223, 2010.

[2] M.H. Hamil, V. Young, J. Boger, and A. Mihailidis, "Development of an automated speech recognition interface for personal emergency response systems," *Journal of NeuroEngineering and Rehabilitation*, vol.6, no.26, pp.1-11, 2009.

[3] DD. Hizer, and A. Hamilton, "Emergency Response Systems: An Overview," *Journal of Applied Gerontology*, vol.2, pp.70-77, 1983.

[4] W.C. Mann, P. Belchior, M. Tomita, and B.J. Kemp, "Use of personal emergency response systems by older individuals with disabilities," *Assistive Technology*, vol.17, pp.82-88, 2005.

[5] E.J. Porter, "Wearing and using personal emergency response systems," *Journal of Gerontological Nursing*, pp.26-33, Oct. 2005.

[6] www.phon.ucl.ac.uk/resource/scribe/scribe-manual.htm, "SCRIBE Manual v.1.0," *University College London, Division of Psychology & Language Sciences, Speech, Hearing & Phonetic Sciences*, 2004.

[7] V. Young, and A. Mihailidis, "Difficulties in Automatic Speech Recognition of dysarthric speakers and the implications for speech-based applications used by the elderly: A literature review," *Assistive Technology Journal*, vol.22, pp.99-112, 2010.