Towards Age-friendly E-commerce Through Crowd-improved Speech Recognition, Multimodal Search, and Personalized Speech Feedback

Lei Meng†, Quy Hy Nguyen*, Xiaohai Tian*, Zhiqi Shen*, Eng Siong Chng†,
Frank Yunqing Guan*, Chunyan Miao†, and Cyril Leung‡

*Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), Nanyang Technological University, Singapore
†School of Computer Science and Engineering, Nanyang Technological University, Singapore
‡Department of Electrical and Computer Engineering, the University of British Columbia, Vancouver, BC, Canada
Email: {lmeng, Nguyen.QH, XHTian, ZQShen, aseschng, yunqing.guan, ASCYMiao, CLeung}@ntu.edu.sg

Abstract—This paper presents an age-friendly system for improving the elderly’s online shopping experience. Different from most related studies focusing on website design and content organization, we propose to integrate three assistive techniques to facilitate the elderly’s browsing of products in E-commerce platforms, including the crowd-improved speech recognition, the multimodal search, and the personalized speech feedback. The first two techniques, namely, the crowd-improved speech recognition and the multimodal search, work together to allow the elderly search for desired products flexibly using either speech, an image, text, or any combination of them whichever are convenient for the elderly. The personalized speech feedback provides a speech summary of search result in a personalized voice. That is, the elderly are allowed to choose or even create their desired voices, and also can customize the voices in terms of pitch, speaking speed, and loudness. As a whole, the proposed system is expected to help and engage the elderly’s E-commerce adoption. Testing on real-world E-commerce product datasets demonstrated the usability of the proposed system.

Index Terms—Age-friendly E-commerce, enhanced user browsing, Crowd-improved speech recognition, Multimodal search, Personalized speech feedback

I. INTRODUCTION

In the past one decade, a dramatic growth in online retailing and electronic commerce (E-commerce) has been observed across the world. E-commerce, defined as the purchase and sale of information, products, and services using any one of the thousands of computer networks that make up the Internet [1], has brought about a revolution in business practices. Through E-commerce, many businesses have been able to increase profitability, gain market share, and improve customer service.

Older people have a great deal to gain from E-commerce, leading to a huge market and business potential with E-commerce for the elderly. According to the World Population aging 2013 report [2], the population aging is taking place in nearly all countries in the world. Globally, the number of older persons aged 60 years and over is expected to more than double by 2050, from 841 million in 2013 to more than 2 billion in 2050. Beyond using the Internet to communicate with family and friends, acquire information, and evaluate services, the senior population has an increasing trend to adopt E-commerce transactions, utilize online access to financial services such as banks and brokerage firms, obtain travel information, and research health-related services.

There are, however, open challenges and issues for the elderly people exploring the E-commerce websites. Some research [3] found that the elderly users aged over 65 years old were over 40% slower than those of the younger generation in using the internet, and were more likely to give up their trials. One explanation is the decline of physical and cognitive functions of the elderly [4]; while another reason is that those aged people are not computer technologies savvy in comparison with the young generation. As such, developing age-friendly user interface and functions to improve the usability of an E-commerce website for the elder users can be a predictor of success for modern E-commerce [5].

Existing research efforts for the age-friendly E-commerce mainly focus on the discovery of reasons for age-related digital divide [6] and various design principles for age-friendly websites to facilitate the elderly users browsing the webpages [7], [8], [9]. The challenges and opportunities with E-commerce for the elderly have also attracted the attention from industry. China’s Baidu launched a portal that caters to the elderly [10]. Besides, Amazon has launched a specific online store targeting at elderly users and their caregivers, namely, the Amazons 50+ Active & Healthy Living†. However, these efforts typically try to improve the elderly’s browsing experience on the Web in terms of simplifying the webpages to facilitate their understanding and gathering and organizing information that are expected to be interesting to them; while limited attention has been given to assistive techniques that help the elderly during online shopping, such as flexible ways for the elderly to search and navigate products, convenient approaches to facilitate the elderly finding desired products, and personalized intelligent assistants that provide online help and suggest products of the elderly’s interests.

†http://www.amazon.com/b?node=5856180011
To address the aforementioned problems, we propose an age-friendly system for improving the elderly’s online shopping experience, empowered with multimodal search and personalized speech feedback capabilities. Specifically, this system includes three functional modules, namely, the crowd-improved speech recognition, the multimodal product search, and the personalized speech feedback. The proposed crowd-improved speech recognition engine (CISR) is a two-step speech recognition model leveraging the Microsoft Speech Platform and the Bing Voice Recognition API, which enables accurate speech recognition in real-time through a personalized and crowd-improved product list; the multimodal product search engine (MM-PSE) is built upon a newly proposed multimodal search algorithm, namely, the online multimodal co-indexing adaptive resonance theory (OMC-ART) [11], which works together with CISR to allow the elderly to search for desired products flexibly using either speech, an image, text, or any combination of them; and the personalized speech feedback (PSF) model makes use of voice conversion techniques [12] to present the summary of search results to the elderly with personalized voices. The elderly is also allowed to create new personalized voices or choose the voice from a list of predefined voices and customize the voices in terms of pitch, speaking speed, and loudness. Prototype systems for these functional modules have been developed and tested with real-world E-commerce transaction data.

The reminder of paper is organized as follows. Section II reviews related work for age-friendly E-commerce. Section III formulates the studied problem. Technical details and implementation of the proposed system are presented in Sections IV and V, respectively. The last section summarizes and highlights future directions.

II. RELATED WORK

With a target of improving the elderly’s online shopping experience, the study of age-friendly E-commerce is related to those on theories and technologies that aim to facilitate the elderly to use E-commerce applications, navigate E-commerce websites, and conduct E-commerce related activities.

A large body of related studies is on Internet adoption by the elderly with a focus on factors resulting in age-related digital divide [13], [14], [6], wherein E-commerce is often treated as a case study. Through user study based approaches, observations and factors ranging from physical behavior to living styles are explored [15], [16], [17], [14], [18], [19], [20], [6]. For example, an early study [15] explores the effect of web page design for the elderly by a user study on two types of websites, in terms of fonts, multimedia content, and diversity of inputs. An eye-tracking based study [16] explores how well the elderly use a search engine. The studies [14], [18] point out that usefulness, ease-of-use, and Internet safety are key indicators for use intention; resource savings and self-actualization as motivators determine senior citizens’ usefulness perception, whereas computer anxiety and computing support are antecedents of ease-of-use.

Another line of research lies on theories and techniques that help web browsing of the elderly, which can be categorized into three key directions. Studies of the first direction aim to explore principles for web page design [17], [7], [8], [9], [21]. In [17], Thirty-six websites designed for older adults were evaluated as to how well they complied to 25 “senior-friendly” guidelines recommended by the National Institute of Aging. “Non Browser” [7] is proposed as a new web design, in which functionality is both restricted and presented differently. The recent study [21] evaluates the usability of E-commerce website for the elderly and highlights the importance of clear and smart navigation design for senior customers. The second direction for age-friendly web browsing [22], [23] explores software enhancements to Web browser to give users access to control their browsing environment. The last direction focuses on the exploration of novel assistive functions. Google investigated and launched the function of voice search [24], [25], [26] and developed the search engine enabled visual search “Google images”. The study [27] presents a user study on using voice demands for the elderly when browsing websites. Assistive virtual agent is proposed in [28] for commerce applications to enhance accessibility for older users, in which the elderly consumers’ preferences for the characteristics of the virtual agents are evaluated, in terms of modality, interaction style, animation, realism, embodiment, and gender.

III. PROBLEM STATEMENT

In view that most related studies focus on theories and design principles of websites for improving the elderly’s E-commerce/web utilization and navigation, while few attempts have been done in terms of innovative assistive techniques, this study aims to develop assistive technologies to enhance the elderly’s browsing experience in E-commerce platforms. Specifically, we propose an age-friendly system with flexible multimodal search and personalized speech feedback for enhanced browsing experience for the elderly. With the elderly’s query in terms of photos, text, speech, and arbitrary combination of them, the proposed system is able to retrieve all similar products in the E-commerce platform at a high accuracy, and a personalized speech feedback is subsequently generated to summarize the search result for the elderly. To fit the speech to different requirements of the elderly, the elderly could be able to tune the styles of the speech, such as speaking speed and volume, and even to create their desired voices.

2https://www.nia.nih.gov/
3https://www.google.com.sg/imghp?hl=en&ei=DiTo_VoOfAo27uASPw7_4D A&ved=0EKouCAIoAQ
IV. ENHANCED USER BROWSING FOR AGE-FRIENDLY E-COMMERCE

A. Overview

Looking through E-commerce websites to search for desired products is a core part in E-commerce; while typing for keyword-based query and the dramatic number of products in terms of brands and categories often bring difficulties for the elderly in browsing products on E-commerce platforms [21]. Towards age-friendly E-commerce, we propose a system with flexible multimodal search and personalized speech feedback to improve the browsing experience of the elderly in E-commerce platforms. As shown in Figure 1, this system includes mainly three functional modules, namely, the crowd-improved speech recognition, the multimodal product search, and the personalized speech feedback. The integrated system enables the elderly to flexibly search for products using a combination of photos, text, and voice as inputs, and employs a personalized speech feedback function that provides a summary of the product search result, such as the number of products and the number and key information of different types of products involved, in a personalized voice, which can be tuned or created by the elderly. In the following sections, technical details of the three modules are presented.

B. Crowd-improved Speech Recognition

The crowd-improved speech recognition (CISR) system enables the voice search function of the system by recognizing users’ voice queries for product search and then feeding the recognized text to the search engine. As illustrated in Figure 2, CISR employs a two-stage model for speech recognition:

- **Stage 1**: with users’ voice queries captured by Web Audio [29] capability of modern web browsers, Microsoft Speech Platform [30] is used to recognize the speech in real-time by a specific grammar built from a personalized and crowd-improved product list.

- **Stage 2**: if the recognition from the first stage yields low-confidence score, i.e. the voice query does not match any grammar in stage 1, Bing Voice Recognition API [31], as a fallback mechanism, is activated to perform speech-to-text recognition.

The Microsoft Speech SDK is chosen from other speech recognition platforms, including Bing Voice Recognition API, Google Speech API, Nuance, and Wit.ai, due to its extensibility to incorporate external grammars and its convenience for implementation on server; while Bing Voice Recognition API is employed mainly because of its reasonable performance and engineering flexibility as a general purpose recognition engine, which can be easily substituted by many others for further development when necessary, such as Google Speech API and Nuance etc.

It is worth mentioning that the grammar-based speech recognition in Stage 1 provides a real-time process of users’ voice inputs, and may accurately recognize keywords, such as product names, from users’ speech with the product list. To optimize the recognition performance gain, the product list could be constructed in two manners:

- **Personalized**: the product list is constructed using current
users’ purchase and browsing history.

- **Crowd-sourced**: the product list is constructed using the recent most popular products in an E-commerce platform.

The personalized product list enables the proposed speech recognition model to react fast and accurately for different users’ speech by learning from individual user’s browsing behavior; while the crowd-sourced product list takes advantage of collective intelligence to make the proposed model context-aware when applied to different E-commerce platforms.

**C. Multimodal Product Search**

The multimodal product search engine (MM-PSE) aims at providing users flexible ways to search for their desired products in the E-commerce platform. MM-PSE is built upon a newly proposed algorithm for the multimodal co-indexing and retrieval of weakly labeled web images, namely, the online multimodal co-indexing adaptive resonance theory (OMC-ART) [11]. As shown in Figure 3, MM-PSE includes a product search module for processing user queries and retrieving relevant products for the user, and a product indexing module for product index base creation and update. We observe that with the indexing and ranking algorithms of OMC-ART, MM-PSE enables users to search for products flexibly using multimodal queries, such as an image, several keywords, or a combination of both, by creating a multimodal index base of E-commerce products using product photos and the associative textual description. In the following two parts, the product indexing and search modules are briefed, of which the technical details can be referred in [11].

1) **Product Indexing**: The utilized co-indexing algorithm of OMC-ART for building the product index base is based on an heterogenous data co-clustering algorithm, i.e. the generalized heterogeneous fusion adaptive resonance theory (GHF-ART) [32]. Given a raw product dataset with product photos and textual description, without additional resources, this algorithm automatically extracts the visual and textual features of products and builds a two-layer index hierarchy, of which the bottom layer contains clusters of similar products while the upper layer contains generalized characteristics of products in the same cluster, represented by cluster weights. Thus, products are co-indexed by the key visual and textual features in this hierarchy, which serves as the product index base and enables image- and/or text-based search. As stated in [11], the proposed co-indexing algorithm is much faster than related state-of-the-arts with light parameter tuning of a ratio value. Besides, the discovered common characteristics of products in the upper layer serves as the basis for the fast and accurate ranking algorithm for product search. More importantly, the built product index base can be online updated, which meets the requirements of E-commerce to handle live big data streams.

2) **Product Search**: The ranking algorithm of OMC-ART takes advantages of the two-layer index hierarchy generated by the co-indexing algorithm for product ranking. Given a query in terms of a photo, textual description, or a combination of both, the ranking algorithm first compares the extracted visual/textual features of the query and key features of the generalized characteristics of clusters in the upper layer of the index hierarchy; products in the cluster more similar to the query are subsequently compared to the query. Compared with commonly-used ranking algorithms, the proposed ranking algorithm of OMC-ART has been demonstrated more effective in search accuracy and more efficient by using key features as intermedium layer for ranking [11]. In addition, with the batch-mode ranking manner, products in clusters less similar to the query are less likely to be the desired ones for users. As such, with a limited number of search result, the ranking algorithm may stop without traversing the entire index base if the ranking list keeps unchanged for a certain period of time, in view that the images most similar to the query are presented prior to those of lower similarity.

**D. Personalized Speech Feedback**

The personalized speech feedback (PSF) engine aims to present the summary of search results with personalized voices, which can either be created by the elderly or chosen from a list of pre-defined voices and can be customized by the elderly.

As shown in Figure 4, once receiving a textual description of the search results, PSF first employs the IVONA text-to-speech (TTS) engine\(^4\) to generate the speech with IVONA’s voice, which subsequently will be transformed to the voice designated by the elderly using voice conversion techniques. This voice can either be chosen from a list of pre-defined voices or created by the elderly through voice recording. Furthermore, the elderly are allowed to make the personalized voices more comfortable by customizing the characteristics of the voices, in terms of pitch, speaking speed, and loudness.

To ensure the high quality of the generated personalized voices, PSF utilizes a newly proposed voice conversion

\(^4https://www.ivona.com/\)
technique [12] to transform the IVONA’s voice into the designated voice. Compared with other voice conversion methods which have been shown to have own advantages and drawbacks [33], [34], the adopted technique uses a system fusion framework to leverage the merits of different conversion methods [35], [36]. Please refer technical details in [12].

It is notable that, besides the voice conversion, there are other methods to generate the voice of a specific person, such as TTS. However, building a personalized TTS system requires the long-time recording of the target speaker and technical skills to manually generate the training data. Therefore, PSF is developed with the cost-effective voice conversion techniques.

V. SYSTEM IMPLEMENTATION

In this section, we illustrate the functionality and user interface of the developed system. At current stage, the multimodal product search engine has been implemented in a Linux environment and tested using real-world ecommerce data, namely, the REC-TMALL dataset\(^5\); while the crowd-improved speech recognition and personalized speech feedback functions are implemented in a Windows environment and tested using a real-world E-commerce product dataset. Note that the integrated system of the three functional modules above has been in development and the following demonstration is based on our tested version.

A. Crowd-improved Speech Recognition

Figure 5 illustrates the user interface of the crowd-improved speech recognition system. The interface provides users with a microphone button to record speech search query directly in the web browser. When the recording function is active, a volume indicator below the search box shows up to help the users make recordings with appropriate loudness. As described in Section IV-B, our proposed CISR system uses a pre-defined grammar in its first stage to improve the recognition. Table I gives examples on the grammar used. In those examples, product titles in the selected product list are used to extract certain keywords such as brand names and product categories. After that, the system automatically creates recognition grammar using certain combinations of those keywords. Note that “*” is the wild card that matches any words so that keywords in grammars appearing in a user’s speech will be accurately detected in Stage 1 of CISR in real-time. In our current implementation, the extraction of keywords from product titles requires human intervention with certain filtering techniques.

<table>
<thead>
<tr>
<th>Product title</th>
<th>Keywords</th>
<th>Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sennheiser RS120 On-Ear 926MHz Wireless RF headphones with Charging Cradle</td>
<td>Sennheiser, headphone</td>
<td>* Sennheiser *.</td>
</tr>
<tr>
<td>Transcend 8 GB Class 10 SDHC flash memory card (TS8GSDHC10E)</td>
<td>Transcend, memory card</td>
<td>* Transcend *.</td>
</tr>
</tbody>
</table>

Table II illustrates two examples on the comparison of speech recognition results from the Bing voice recognition engine, the Stage 1 of CISR, and the output of CISR. The first example demonstrates the behaviors of different systems with an input speech when the requested product (i.e. Sennheiser headphone) is inside the product list for CISR, while the second example demonstrates those when

<table>
<thead>
<tr>
<th>Input speech</th>
<th>Recognition engine</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do you have any Sennheiser headphone?</td>
<td>Bing voice recognition engine</td>
<td>Do you have anything either headphone?</td>
</tr>
<tr>
<td>Stage 1 of CISR</td>
<td>Sennheiser headphone</td>
<td></td>
</tr>
<tr>
<td>CISR</td>
<td>Sennheiser headphone</td>
<td></td>
</tr>
<tr>
<td>I want to buy a Samsung Chromebook.</td>
<td>Bing voice recognition engine</td>
<td>I want to buy a Samsung chromebook.</td>
</tr>
<tr>
<td>Stage 1 of CISR</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>CISR</td>
<td>I want to buy a Samsung chromebook.</td>
<td></td>
</tr>
</tbody>
</table>

\(^5\)https://tianchi.aliyun.com/datalab/index.htm?spm=5176.100075.2368.2.5c zf7X
the requested product (i.e. Samsung Chromebook) is outside the product list. We observe that, as a general purpose recognition engine without the help of product list, the Bing voice recognition engine may not recognize complicated brand name (in this case, the Sennheiser) correctly. In contrast, the stage 1 of CISR is able to successfully recognize the product keywords. In this case, the Stage 2 of CISR is skipped due to high confidence of the result from Stage 1 so that the final result of CISR is the same as the result of Stage 1. When the requested product is outside the product list, as shown in the second example, the Stage 1 of CISR cannot produce any results with high confidence. However, CISR still obtains high-quality result by activating the Stage 2, which uses a general purpose recognition engine, currently the Bing voice recognition engine, without constraint of the product list grammar. In this way, CISR ensures a robust performance in speech recognition through the two-stage model.

B. Multimodal Product Search Engine

Figure 6 illustrates the user interface and the functionality of the developed multimodal product search engine using an example of search results on the REC-TMALL dataset. As observed, the designed user interface follows that adopted by traditional ecommerce platforms with a “camera” icon to upload images, which will be displayed on the left side of the search bar.

As shown in Figure 6a, by uploading an image without text, the search engine will return a list of products with similar product photos in a top-down left-right order in terms of similarity. This similarity is, as suggested in [11], measured by three types of visual features, including the grid color moment, the edge direction histogram, and the wavelet texture. We observe that three products in the top four returned ones are of the same brand to the uploaded image. Note that the uploaded image is exactly that of the top one returned product. The cloth of the top third product is not of the same style to the query though, it is visually similar in terms of the whole. Because all related products are listed in the top four products, the rest products are not similar to the query. However, these products are also related to fashion cloth. These findings demonstrated the effectiveness of the search engine.

As a flexible multimodal search engine, besides visual search, users are also allowed to query by text or a combination of both image and text, as shown in Figures 6b and 6c. Note that the text is in Chinese because the utilized REC-TMALL dataset is from a Chinese E-commerce platform. From Figure 6c, we observe that, with additional textual description in the query, the search result has a significant improvement, compared with those in Figure 6a; besides the products of the same brand to the query, the rest retrieved products are all about fashion lady cloth of Korean style, which meet all requirements in the textual description. We also observe that the search result using solely keywords in Figure 6b is quite different from that in Figures 6a and 6c. This is because the multimodal search engine, namely, the OMC-ART [11], is able to make a tradeoff for measuring the similarity between the query and products in the index base, which is an automatically leaned parameter during the product indexing process and can be fine-tuned manually.

C. Personalized Speech Feedback

As described in Section IV-D, PSF allows the elderly to select preferred voices, create personalized voices, and customize voices. Figure 7 shows the user interface of PSF. As observed, the pre-defined and personalized voices are listed in different sectors, and the elderly may play the audio sample of each voice by clicking the flag on the right side of the speaker’s name, say “amy”, and select their favourite ones for speech
Fig. 7: The user interface of the prototype system with the personalized speech feedback engine.

Please read this sentence: Many farms could be bankrupted as a result.

Fig. 8: The user interface of the prototype webpage for voice recording.

Please read this sentence:

Many farms could be bankrupted as a result.

Fig. 9: The pop-up panel for voice customization.

“Start Voice Conversion Training” button at the bottom of the webpage (Figure 8b) to train a personalized transformation function. When completed, the trained voice will be added to the user interface for voice selection, in our case shown in Figure 7, the option “Simon” under the “Personalized voices” list. It is worth mentioning that to achieve a robust transformation function, the recording should be done in a quiet environment, and the speaker is preferred to use a neutral speaking style.

Beyond the creation of personalized voices, the elderly can click the “Tune voice” button to modify any voices in the lists of predefined voices and personalized voices through a pop-up panel. As shown in Figure 9, the elderly may tune the characteristics of the selected voice to customize the voices to be more comfortable to them, in terms of the speaking speed, the pitch, and the loudness. Besides, a preview function is provided for checking the customized voices. After that, user can save the changes by clicking the “Save Configuration” button.

VI. CONCLUSION AND FUTURE WORKS

This paper presents a system for age-friendly E-commerce, targeting at enhancing the elderly’s browsing experience in E-commerce platforms. Different from most existing work focusing on age-friendly website design, the proposed system takes advantage of recent advanced search and speech technologies, and provides assistive functions for the elderly, including crowd-improved speech recognition, flexible multimodal search, and personalized speech feedback. The crowd-improved speech recognition function makes use of a crowdsourced and personalized product list to provide fast and accurate recognition of the elderly’s voice query. The flexible multimodal search engine, working together with the crowd-improved speech recognition function, support speech, visual, and text queries for product search, in terms of individual or any combination of them whichever are convenient for the elderly. The personalized speech feedback function provides a speech summary of search result in a personalized voice for the elderly, which is aimed to facilitate the elderly to understand the large amount of returned products relevant to their query.
Note that through the personalized speech feedback function, the elderly are able to customize the synthetic voice, such as volume and speed, and even create their preferred voice.

Despite the current achievement, our future work mainly includes three aspects. First, the three testified functions of the proposed system will be implemented into one platform. Second, a series of user study will be conducted to evaluate the system and guide further research on system improvement. Third, other assistive techniques will be explored to enrich the system, such as age-friendly online assistant and personalized product recommendation.

ACKNOWLEDGEMENT

This research is supported by the National Research Foundation, Prime Minister’s Office, Singapore under its IDM Futures Funding Initiative.

REFERENCES


