Design and Implementation of a Disambiguity Framework for Smart Voice Controlled Devices

Kehua Lei[1†], **Tianyi Ma**[1†], **Jia Jia**[1*], **Cunjun Zhang**[1] and **Zhihan Yang**[1]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Abstract

With about 100 million people using it recently, SVCD(Smart Voice Controlled Device) are becoming demotic. Whether at home or in an office, usually, multiple appliances are under the control of a single SVCD and several people may manipulate an SVCD simultaneously. However, present SVCD fails to handle them appropriately. In this paper, we propose a novel framework for SVCD to eliminate orders' ambiguity for single user or multi-user. We also design an algorithm combining Word2Vec and emotion detection for the device to wipe off ambiguity. Finally, we apply our framework into a virtual smart home scene and the performance of it indicates that our strategy resolves the problems commendably.

1 Introduction

Smart Voice Controlled Device(SVCD) is prevalent around the world. Among U.S. adults, 53 million people now own at least one SVCD. The total number of devices in homes increases by 78 percent per year. Despite its growing popularity, the capability of SVCD remains to be enhancing. Some existing researches aspire to explore problems and put forward design guidelines for the devices[Luger and Sellen, 2016][Porcheron *et al.*, 2018][Myers *et al.*, 2018]. Others struggled to find new interaction techniques[Corbett and Weber, 2016][Lyons *et al.*, 2016][Zhong *et al.*, 2014] and algorithms[Rong *et al.*, 2017] to enhance the SVCD system's efficiency. However, very few studies propose a specific method to improve user experiences. In this paper, we focus on the two most outstanding problems within the SVCD. The first problem is about the ambiguity elimination in multi-device situation. It is commonplace that the user manipulates SVCD with a sentence that is semantically-ambiguous, which often cause a dilemma for SVCD. The second problem is that when multiple users manipulate the SVCD synchronously, the current devices usually fail to make an eclectic decision that meets everybody's demand.

[1*]Corresponding author: Jia Jia (jjia@mail.tsinghua.edu.cn)

[2†]Joint first author: Kehua Lei, Tianyi Ma (According to the alphabetical order of first name)

To address the issues, we propose a novel framework for SVCDs and design a precise algorithm for them. Further, we implement this new framework in a real smart home scene, which receives a satisfying result. In our framework, the SVCD first uncouples the acoustic commands from different people. Then it eliminates the ambiguity in each user's instruction with the algorithm. Following that, the SVCD shall decide the priority based on the acoustic information and the level of ambiguity of their commands. If the device can decide the priority, it will execute an instruction or start a multi-round dialogue with the top priority person. Otherwise, the device guides the users to make a compromise. As for the algorithm, an SVCD possesses a depository storing abundant default instructions. The device calculates the semantic distance between the users' commands and the default instructions. Based on the distance and the user's emotion and habit, the SVCD surmises the user's intent. If ambiguity still exists, the device shall embark on a multi-round dialogue with the user. We implement our SVCD in a virtual smart home and invite several testers to compare our SVCD with a current SVCD XiaoMi smart speaker. Our device receives an average score of 7.14 out of 10 while the other device get an average of 5 out of 10, which indicates that our framework and algorithm is effective.

Our contribution can be summarized as follow :

- We discover that recent SVCDs are incapable of handling ambiguous command in multi-device scene and dealing with commands made by several users simultaneously. Thus, we propose a novel framework to resolve the problems and achieve a satisfying result.
- We design an algorithm to eliminate ambiguity in the user's instruction. Mapping the user instruction to default instructions, the SVCD can eliminate some ambiguity. Assisting by user's habit and emotion, the device wipes off more ambiguity. If ambiguity still exists, the SVCD will embark on a multi-round dialogue with the user to make a decision.

2 Framework and Implementation

2.1 Basic Framework

Figure 1 shows the framework and basic algorithm we proposed. The framework can be dissected into three parts: default setting, ambiguity elimination, and priority-deciding.

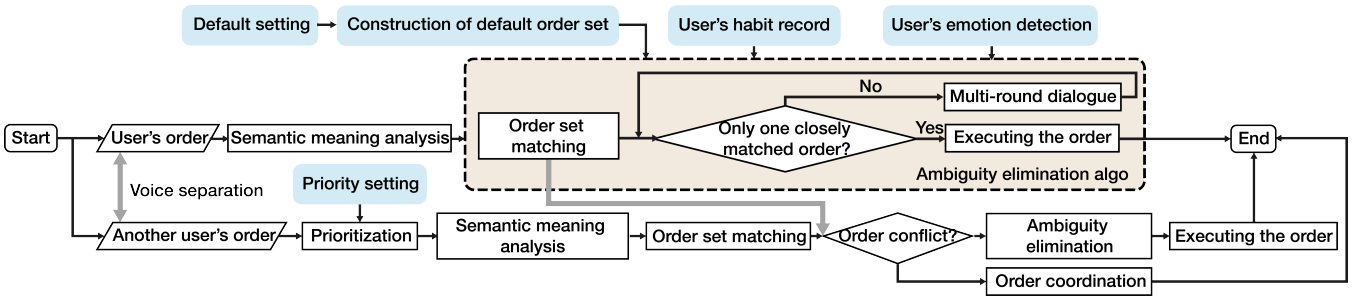

Figure 1: Framework

The prerequisite for users to manipulate SVCD is to do default setting, such as setting the location of the device. The SVCD generates the default instruction set based on the setting. For each user's command, the SVCD applies the algorithm to wipe off ambiguity. Then, based on the acoustic feature of the commands, the device decides the priority. If the attempt failed, the SVCD will detect if semantic conflicts exist within the commands. If yes, the device shall guide the users to reach a consensus. Rather, it shall execute all of them. If the ambiguity exists concurrently, the SVCD will manage to find the intersecting instructions of all the commands and execute them or do further ambiguity elimination.

2.2 Algorithm

The ambiguity-elimination algorithm shall be divided into three part: instruction set matching, emotion and habit assistance, and multi-round dialogue. First, to match the user's command with default instructions, applying a Word2Vec network[Zhang *et al.*, 2015], the SVCD converts the user's and default instructions into semantic vectors. Then, it sorted the default commands by their Euclidean distances to the user's command. If one instruction's distance substantially smaller than others', the device shall surmise it meets user's demand and executes it on hardware. Conversely, if there are only several matched instructions whose distance values are close or no matched instruction exists, it shall detect the user's emotion acoustically by a multi-path neural network[Zhou *et al.*, 2018]. Compressed and mapped, the emotion vector help to pad significant verbal components missed in user's command. Also, with the habit information collected from the user in daily use, the SVCD can conjure the best possible instruction. However, those attempts should fail sometimes. In this case, the device shall embark on a multi-round dialogue with the user to obtain more information.

2.3 Implementation

We implement our framework in a smart home scene. The demonstration, in which we constructed an SVCD with our framework, is a 2.5D virtual smart home space. The space represents the real smart home scene, contains an SVCD and multiple lights that can be manipulated by the user through the virtual smart speaker. Once it receives commands from the microphone, the SVCD will respond immediately. If it succeeds to surmise the user's intent, a light shall be ignited. Conversely, if more information is required, it shall initiate a dialogue with the user automatically.

Figure 2: Implementation Interface

2.4 Experiment

Having implemented our framework in a virtual smart home, we test our SVCD and the result proves that our device constructed with our framework can better meet the demand of the users. We randomly select 10 people(5 males, 5 females) to participate in our experiment. After using a current product XiaoMi smart speaker, they manipulate the light through our SVCD and rate the two SVCDs. Our SVCD receives an average score of 7.14 out of 10 while XiaoMi only receives an average score of 5 out of 10. This result indicates that our framework and algorithm can create more user-friendly SVCDs.

3 Conclusion

In this paper, we discover that present SVCD are incapable of dealing with ambiguous instruction in multi-device scene and commands made by several people simultaneously. To solve the problem, we propose a novel framework and algorithm for SVCDs and implement them in a virtual smart home scene. We invite tester to compare our SVCD with another device. The result demonstrates the effectiveness of our framework and algorithm. Video of our system and related simulations can be found at https://drive.google.com/file/d/1zNKuOnHGZJ_f70HFEj14HWTpokeBQ36d.

Acknowledgments

This work is supported by National Key Research and Development Plan (2016YFB1001200),the Innovation Method Fund of China (2016IM010200), the state key program of the National Natural Science Foundation of China (NSFC) (61831022),and National Natural, and Science Foundation of China (61521002).

References

[Corbett and Weber, 2016] Eric Corbett and Astrid Weber. What can i say?: addressing user experience challenges of a mobile voice user interface for accessibility. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 72–82. ACM, 2016.

[Luger and Sellen, 2016] Ewa Luger and Abigail Sellen. Like having a really bad pa: the gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5286–5297. ACM, 2016.

[Lyons *et al.*, 2016] Gabriel Lyons, Vinh Tran, Carsten Binnig, Ugur Cetintemel, and Tim Kraska. Making the case for query-by-voice with echoquery. In *Proceedings of the 2016 International Conference on Management of Data*, pages 2129–2132. ACM, 2016.

[Myers *et al.*, 2018] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 6. ACM, 2018.

[Porcheron *et al.*, 2018] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*, page 640. ACM, 2018.

[Rong *et al.*, 2017] Xin Rong, Adam Fourney, Robin N Brewer, Meredith Ringel Morris, and Paul N Bennett. Managing uncertainty in time expressions for virtual assistants. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 568–579. ACM, 2017.

[Zhang *et al.*, 2015] Dongwen Zhang, Hua Xu, Zengcai Su, and Yunfeng Xu. Chinese comments sentiment classification based on word2vec and svmperf. *Expert Systems with Applications*, 42(4):1857–1863, 2015.

[Zhong *et al.*, 2014] Yu Zhong, TV Raman, Casey Burkhardt, Fadi Biadsy, and Jeffrey P Bigham. Justspeak: enabling universal voice control on android. In *Proceedings of the 11th Web for All Conference*, page 36. ACM, 2014.

[Zhou *et al.*, 2018] Suping Zhou, Jia Jia, Qi Wang, Yufei Dong, Yufeng Yin, and Kehua Lei. Inferring emotion from conversational voice data: A semi-supervised multi-path generative neural network approach. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.