Designing Multimodal Acoustic Environment Corpus to Improve Speech Interaction in Living Room

Kenichi Shibata [1], *Kengo Ikeya*[1], *Yuki Deguchi*[1], *Yoichi Takebayashi*[1], *Shigeyoshi Kitazawa*[1], *Shinya Kiriyama*[1]

[1] Shizuoka University, Japan

shibata@kitazawalab.net

Abstract

We constructed a multimodal acoustic environment corpus for interaction design. The intelligent environment consists of a living room interacting with different users and adapting it to their preferences. We developed a speech interaction system to control electrical appliances in the living room, and to record and accumulate real-life interaction data to the corpus. The results of interaction analysis indicated that the constructed corpus was effective in investigating the actual conditions of real-life interaction using speech interfaces.

Index Terms: multimodal acoustic environment corpus, speech interaction design, situation understanding, user environment adaptation

1. Introduction

The progress of the consumer electronics (CE) technologies has brought us a secure and comfortable environment in the living room. Many works have done to realize intelligent services considering user environment situations; a user interface agent that provides intelligent context-sensitive help and assistance for a network of consumer devices [1], designing an intelligent living-space with automatic control system to control all home appliances in the space [2], a study of developing a voice recognition system for reaction to human input in assistive environment [3], a detection method of high-level behavior with a model for determining the threshold value dynamically according to individual behavioral pattern [4].

As we spend a lot of time at home, more and more people are looking forward to living comfortably and request different devices with a specific set of functions. Our living rooms, a very important room in the house, are filled with a variety of devices and furniture where each one of us has a favorite arrangement and layout. From this viewpoint, we developed speech interfaces adapting the environment to the diversity of users' vocabularies, room arrangements, or required function of the CE devices [5]. This paper describes the multimodal acoustic environment corpus-based methodology for speech interaction design in a living room where the environment adapts itself to user's preferences (see Figure 1).

2. Speech Interaction System based on Multimodal Acoustic Environment Corpus

In order to understand the peculiar situation of each user of the environment and provide him/her with adaptive services or functions via the speech interaction system, it is important to analyze real-life interaction data from various perspectives. We have already established a human behavior analysis environment based on a multimodal behavior corpus as shown in Figure 2 [6]. The corpus consists of videos and speech data

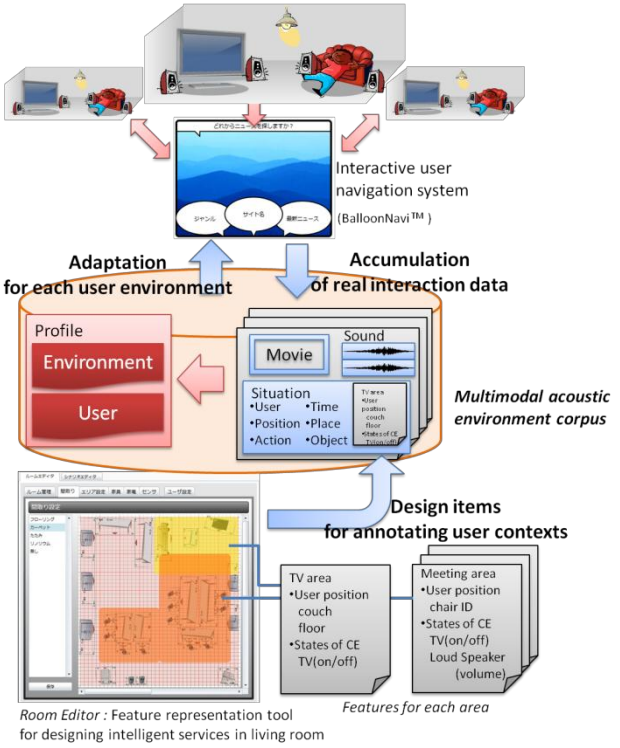

Figure 1: *A corpus-based speech interaction system*.

of real-life human interactive behavior and appended annotations from multiple viewpoints. The most significant point is that the annotations of feature representation are organized in a structure that is flexibly designable according to analysis purposes. The structured annotation data helps us to conduct corpus analyses, since we can easily retrieve and compare the examples of the same features from various viewpoints. Inheriting this idea, we constructed a multimodal acoustic environment corpus and designed an intelligent speech interaction system by adapting speech recognition errors or behavioral styles for each user and acoustic features for each room.

To represent essential aspects of each example in the corpus, we developed a software called Room Editor, a tool for creating service-oriented room models. Room Editor provides functions not only to describe the layouts of living rooms, the arrangements of furniture and equipment, but also to give names to functionally meaningful areas and items and to describe features for each area. The created room models are used to describe features for each situation on the corpus.

We also developed a speech interaction system for accumulating corpus interaction data and providing users with adaptive services and functions by using profiling information for each user environment from the corpus. To control the

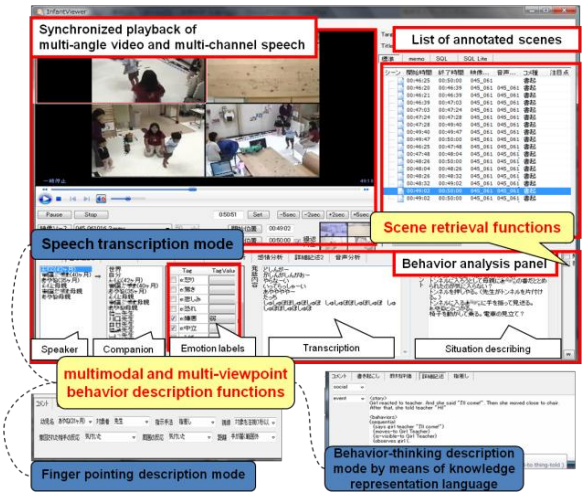

Figure 2: *A multimodal human behavior corpus*.

interaction between users and the system, we introduced BalloonNavi ™[7], an interactive user navigation system. The system allows users to navigate to the information they require from their interaction history. We integrated the speech input function into the system by using the Open-Source Speech Recognition Engine Julius [8].

3. Corpus Analysis for Speech Interaction Design

We conducted an experiment for studying how to utilize the corpus for designing adaptive speech interaction. A simple scenario to customize preferred Internet news site was prepared and five subjects participated in the evaluation. All sessions were recorded in video and audio, and the speech data was segmented into utterances and was given annotation data based on the feature description structure designed by the room model. Speech recognition errors were semi-automatically accumulated to the corpus.

To design speech interaction systems, the problems of recognition errors are unavoidable. In order to study the actual conditions of recognition errors and consider solutions for each case, we analyzed the corpus by focusing on examples of speech recognition errors.

Figure 3 shows an instance of the analysis. In this case, the utterances of substitution errors are retrieved and listed in frame *A*. The corpus is equipped with rich and structured annotation data that allows us to retrieve various conditions such as the speakers' IDs, positions, or distances from the microphones. So this list is just an example. If we focus on the substitution error list example, we can see the whole session of the interaction data in frame *B*. From this we learned the system failed twice to recognize the speech but worked well the third time. What was the difference between the second and third time? Two hypotheses were generated by checking the video; one was the direction of the user's head and his distance from the microphone. The other was the used vocabulary. The second utterance was /s a i t o m e: / which means "the name of site" while the third one was /s a i t o/ which only means "site". Seeing the waveforms for the two utterances indicated that the S/N level of both was the same. Consequently, the difference of the sound source had less influence than the user's vocabulary in this situation. The analysis results using the corpus gave us fruitful findings to understand acoustical situations surrounding users in speech interactions with the system.

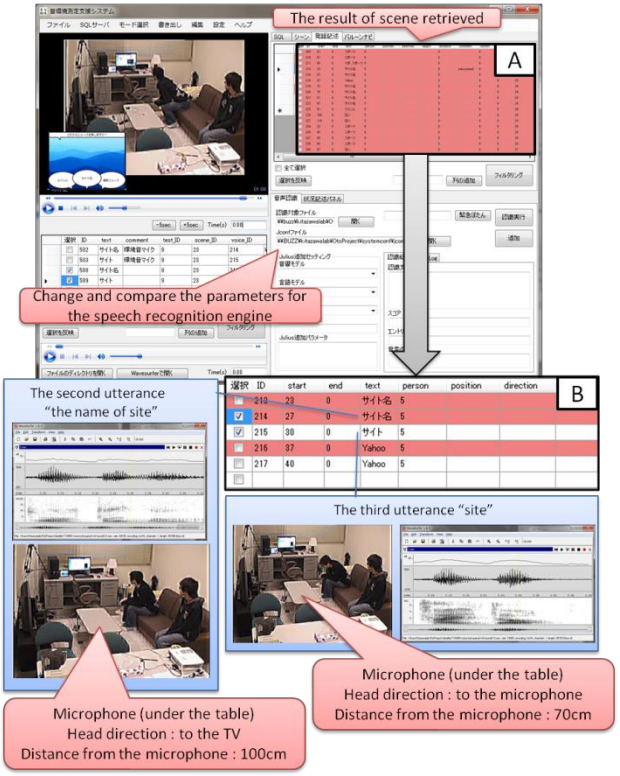

Figure 3: *An analysis of speech recognition errors using the multimodal acoustic environment corpus*.

4. Conclusions

We have built a foundation for speech interaction design based on a multimodal acoustic environment corpus. The corpus has structured annotation data, which facilitates interaction analyses from multiple viewpoints. The experimental results proved that the constructed corpus was effective in investigating the actual conditions of real-life speech interaction with the user navigation system and in obtaining valuable findings for improving speech interfaces.

5. References

[1] Henry Lieberman et al, "A goal-oriented interface to consumer electronics using planning and commonsense reasoning," Knowledge-Based Systems, Vol.20, Issue 6, 592-606, 2007.

[2] Chun-Liang Hsu et al, "Constructing Intelligent Living Space Controlling System with Bluetooth and Speech-Recognition Microprocessor," Eighth International Conference on Intelligent Systems Design and Applications, 2, 666-671, 2008.

[3] Eric Becker et al, "Event-based Experiments in an Assistive Environment using Wireless Sensor Networks and Voice Recognition," PETRA'09, 2009.

[4] H. Yamahara et al, "Behavior Detection Based on Touched Objects with Dynamic Threshold Determination Model," EuroSSC2007, Lecture Notes in Computer Science, 4793, 142-158, 2007.

[5] K. Shibata et al, " A Study of Speech Interface for Living Space Adapting to User Environment by Considering Scenery Situation," The 9th International Conference on Auditory-Visual Speech Processing, 186-189, 2010.

[6] S. Kiriyama, et al, "Mental-State Analysis for Understanding Children's Behavior Based on Emotion-Label Sequences in Multimodal Speech-Behavior Corpus," Oriental-COCOSDA, 2010.

[7] Details about BalloonNavi™ (Japanese only) http://balloonnavi.digital-sensation.jp/

[8] A. Lee et al, "Recent Development of Open-Source Speech Recognition Engine Julius," APSIPA ASC, 2009.