Intelligent User Interface Based on Multimodal Dialog Control for Audio-visual Systems

Hiroshi Shinjo
Ui Yamaguchi
Akio Amano
Konagi Uchibe
Atsushi Ishibashi
Hideki Kuwamoto

OVERVIEW: As the functionality of AV equipment improves and the number of TV broadcast channels increases, the era in which we "can't pick and can't watch" our preferred programs is approaching. In addressing this problem, Hitachi has developed an "intelligent user interface"—which uses "multimodal dialog control" to help users (i.e. viewers) operate home AV equipment by means of spoken dialog. Integrating speech information processing, image information processing, and a TV program recommender based on viewing-history analysis, this multimodal dialog control can partake in dialog with the viewer. Without the need for keyword entry, the TV program recommender automatically analyzes the viewer's program preferences from their past viewing history, and then selects, recommends, and records certain programs according to the analysis results. Utilizing voice information processing, this intelligent user interface can operate AV devices via spoken dialog very close to everyday conversation. It can also distinguish different viewers by facial image recognition and automatically change the offered services (such as the TV program recommender) to match the preferences of whoever is in front of the TV.

INTRODUCTION

ACCOMPANYING the start of terrestrial digital broadcasting, the popularization of AV (audio-visual) PCs and HDD/DVD (hard-disk drive/digital versatile disc) recorders is continuing. Moreover, from now onwards, the storage capacity of hard disks will continue to grow ever bigger, meaning that we are close to the era in which "anything can be recorded." Furthermore, with the connection of various home AV appliances as a network, the functionality and performance of such equipment is improving all the time. Under these circumstances, the utilization pattern of AV equipment, especially TVs, will change from a state of "autonomously watching and recording" favorite TV programs to a state of "can't pick and can't watch."

In response to these challenges, Hitachi has developed a prototype "intelligent user interface" for supporting the operation of AV equipment by means of multimodal dialog control—which integrates

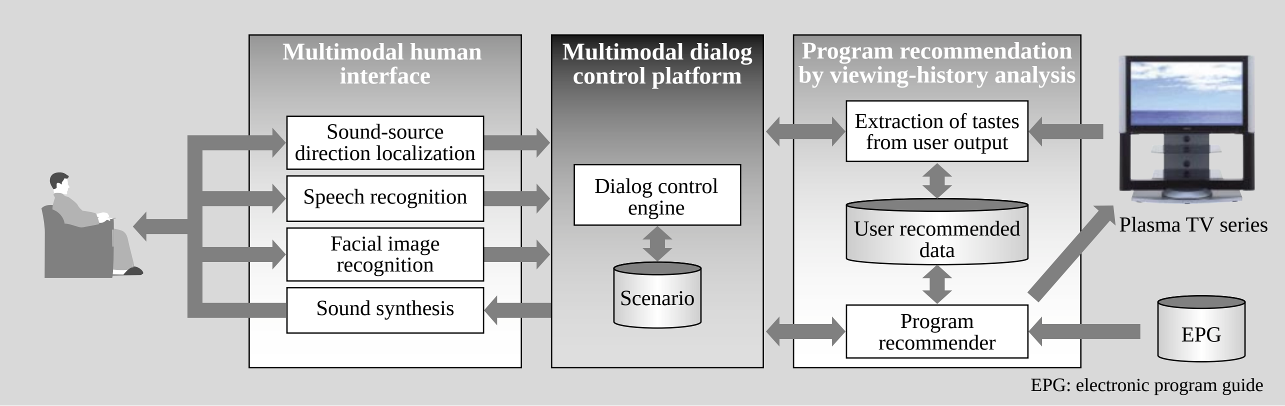
Multimodal human interface
Sound-source direction localization
Speech recognition
Facial image recognition
Sound synthesis
Multimodal dialog control platform
Dialog control engine
Scenario
Program recommendation by viewing-history analysis
Extraction of tastes from user output
User recommended data
Program recommender
Plasma TV series
EPG
EPG: electronic program guide

Fig. 1—Structure of Intelligent User Interface Using Multimodal Dialog Processing.
Based on multimodal dialogs combining various input and output methods (such as sound and pictures), services can be offered in response to user circumstances.

technologies for speech synthesis and recognition, image recognition, and text analysis.

In the rest of this paper, the concept behind this intelligent user interface is explained, and its elemental technologies are introduced.

CONCEPT BEHIND INTELLIGENT USER INTERFACE

Through intuitive operation based on a dialog format used in natural language, the intelligent user interface efficiently selects programs that the viewer might want to watch from a multitude of broadcast and recorded programs, thereby supporting the operation of AV equipment. The interface is constructed as a system that integrates several elemental technologies, namely, voice recognition, facial image recognition, a "TV show recommender" (which analyzes a user's viewing history), and multimodal dialog processing (see Fig.1). The main functions realized by these technologies are summarized as follows:

(1) User-friendly interface

Multimodal dialog processing — which integrates several input mechanisms such as voice recognition, voice synthesis, and image recognition — realizes a user interface that allows human-interface interaction by the same dialog as used in human-human interaction.

(2) TV program recommender

Without the need for inputting any keywords, this function learns the user's program preferences from the viewing history of that user, and recommends and automatically records certain TV shows according to the analysis results.

(3) Provision of services tailored to individual users

The identity of the user in front of the AV device is automatically determined from the user's face; services tailored to that user are then provided. For example, on uttering their name, an individual user is recommended recorded or currently on-air programs according to their preferences by means of voice interaction. Moreover, at the time when a regularly watched program is about to be aired, the viewer is visually and audibly informed.

(4) Monitoring user status

The status of the user is monitored by using a camera and an array of microphones. An example of this function is automatic recording when the viewer is absent and replaying the recorded program when the viewer comes back. When the viewer rises from his/her seat while watching a TV program, the program on-air is interrupted, and recording of that program starts. After that, when the viewer returns to his/her seat, on identification of that person, the program is restarted from the interruption point.

In addition to the above-described functions, expansion aimed at provision of various other services is possible.

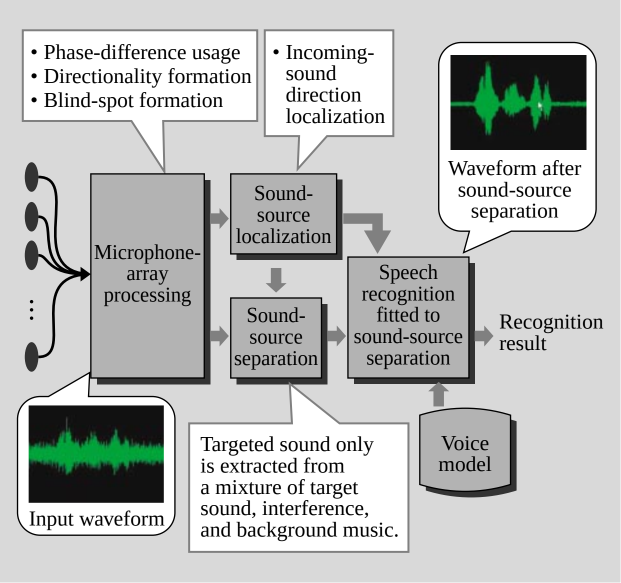
• Phase-difference usage
• Directionality formation
• Blind-spot formation
• Incoming-sound direction localization
Waveform after sound-source separation
Microphone-array processing
Sound-source localization
Speech recognition fitted to sound-source separation
Sound-source separation
Recognition result
Input waveform
Targeted sound only is extracted from a mixture of target sound, interference, and background music.
Voice model

Fig. 2—Flow of Speech-recognition Processing.
Phase and amplitude differences arriving at the microphone array are used to estimate the sound-source direction. The focal point of the target sound is picked out, and speech-recognition processing is carried out on the extracted voice.

TECHNICAL FACTORS

Speech-recognition Technology

By means of a microphone array of several microphones, two functions are realized: estimation of the arrival directions of sound and voice, and sound-source separation that focuses on only a targeted voice under the presence of background noise. By utilizing eight microphones arranged in a two-dimensional array, these two functions can simultaneously detect angle of direction (horizontal direction) and angle of elevation (vertical direction), and separate the source of a particular sound.

As regards sound-source separation, a way of controlling the directionality of an adaptive filter and null beamforming—based on information on the targeted sound and background noise (which both depend on segregation of frequency components)—has been devised[(1)], and background-noise is suppressed to a maximum value of 20 dB (see Fig. 2).

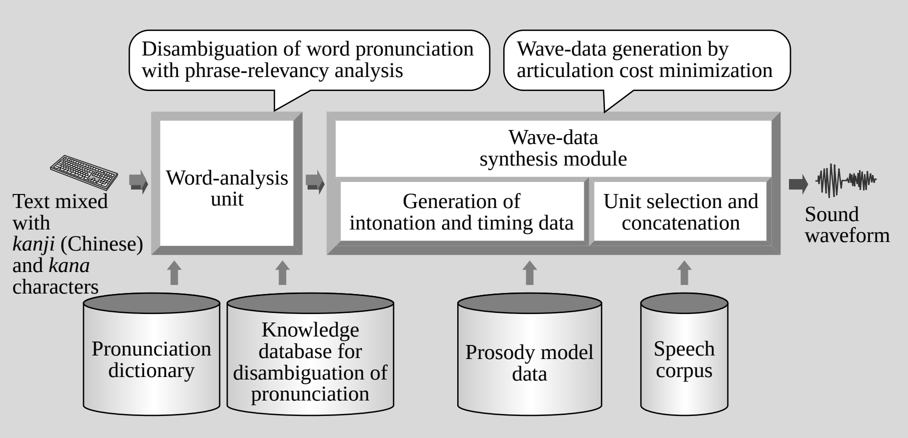
Disambiguation of word pronunciation with phrase-relevancy analysis
Wave-data generation by articulation cost minimization
Text mixed with *kanji* (Chinese) and *kana* characters
Word-analysis unit
Wave-data synthesis module
Generation of intonation and timing data
Unit selection and concatenation
Sound waveform
Pronunciation dictionary
Knowledge database for disambiguation of pronunciation
Prosody model data
Speech corpus

Fig. 3—Flow of Text to Speech Synthesis Combination Processing.
To deal with a text that is mixture of kanji *(Chinese) and* kana *characters, this function is composed of "word processing" (for determining reading and accent) and "waveform synthesis processing" (for generating a waveform by means of selecting an appropriate speech unit from a speech corpus).*

The voice waveform obtained from the sound-source separation is input into a voice-recognition engine[2] (developed for car-navigation use), and voice commands for use with AV equipment (starting with TV channel selection, etc.) are recognized.

Speech-synthesis Technology

A high-grade speech-synthesis technology—which can synthesize a sound giving a high impression of a human voice—was developed. As regards the developed user interface for AV equipment, to precisely utter words that cannot be recorded in advance, replaying recorded speech is not possible, so a method of speech synthesis is needed.

At Hitachi, we have developed our own "articulation cost minimization method"[3]. This method can efficiently select optimum speech units and, by smoothly connecting the selected speech units, it can synthesize a multitude of different sounds with a remarkably high and human-voice impression (see Fig. 3). Moreover, to guide the viewer around the program information by voice, we have developed a technology called "disambiguation of word pronunciation with phrase relevancy analysis" for correctly reading out data aloud from an EPG (electronic program guide). Even in the case of many pronunciations of certain words, the reading can be judged from the context before and after the word in question and from relevant data between words. In this way, intelligent separation of phrases sprinkled with *kanji* (Chinese) and *kana* characters is realized.

Facial Image-recognition Processing

We have developed a technology that extracts the facial image of a user from a screen image in order to identify that person. The processing procedure is summarized as follows:

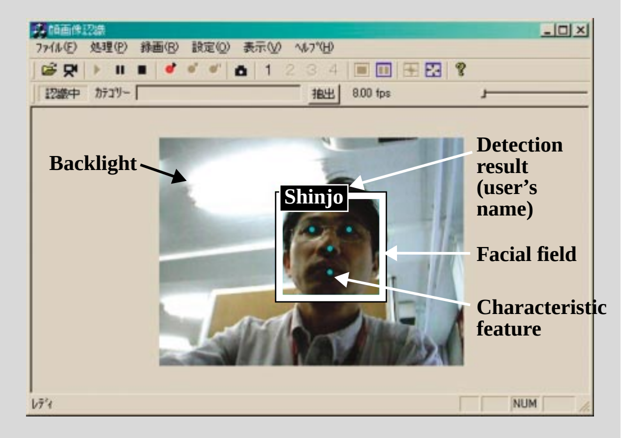

Fig. 4—Example of Face Recognition.
Under any lighting conditions, facial region and features (characteristic points) are extracted in order to distinguish different people.

First, a face-extraction filter is used to detect the facial region in the screen image. After that, the user's facial features (i.e. characteristic points) such as eyes, nose, and mouth are detected, and a "characteristics measure" that expresses individual differences in faces is worked out from the gray image around these features. By comparing this measure with values already stored in a database, it is possible to determine the identity of the user (see Fig. 4).

One problem with this face-recognition technology is how to accommodate movement of set-up location and differences in time of use, fluctuation in lighting conditions such as backlighting and oblique lighting, and changes in facial orientation angle accompanying differences in users' vision. To address these problems, we use a numerical model such as canonical discriminant analysis to improve recognition accuracy.

The facial-image-recognition processing method

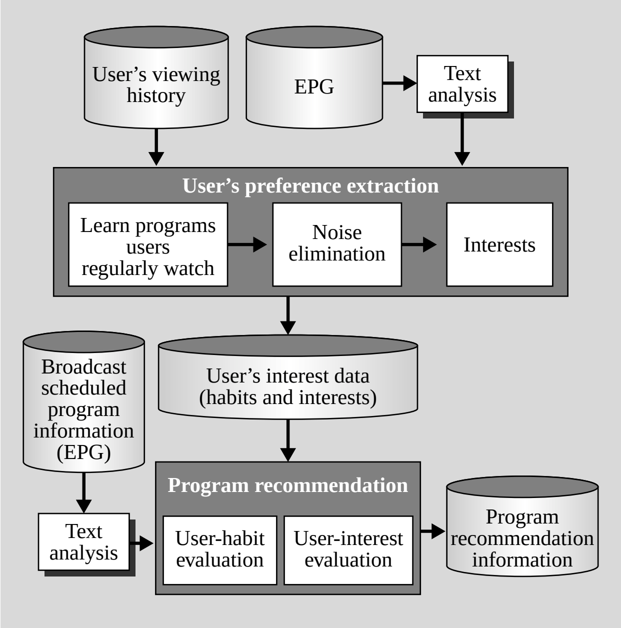
User's viewing history
EPG
Text analysis
User's preference extraction
Learn programs users regularly watch
Noise elimination
Interests
Broadcast scheduled program information (EPG)
User's interest data (habits and interests)
Program recommendation
Text analysis
User-habit evaluation
User-interest evaluation
Program recommendation information

Fig. 5—Overview of Preference-extraction and TV-program-recommender Functions.
Preferences and viewing habits are learned according to the user's viewing history, and TV programs are recommended according to the learning results.

described here was evaluated by using a database containing facial images of 80 people (8000 pictures). The lighting direction in regards to the user's face was –45 to +45 degrees horizontal and 0 to 60 degrees vertical. The facial orientation in both the horizontal and vertical planes was –10 to +10 degrees. The discrimination success ratio for the database used was 98.0% in the case of a front-facing face (not inclined), 95.1% for a face inclined at ±10° in the horizontal direction, and 96.9% for a face inclined at ±10° in the vertical direction.

Program-recommendation Technology by Analyzing History of Past Viewing

User preferences for different program content (i.e. interests) and their daily and monthly viewing habits are learnt, and a "TV program recommender" that reflects the learning results is executed (see Fig. 5). Text such as program name, genre, cast, and briefing sheets from an EPG is analyzed, and groups of keywords are input into the profiler. Viewing history of viewing time and frequency regarding programs watched by the user is also input. On preference extraction, the problem of precision reduction arises because programs such as news ("viewing noise") are watched and learned by the program-recommendation technology regardless of content. As a consequence, viewing noise is eliminated according to the results of viewing habits in order that the user's interests are accurately learnt.

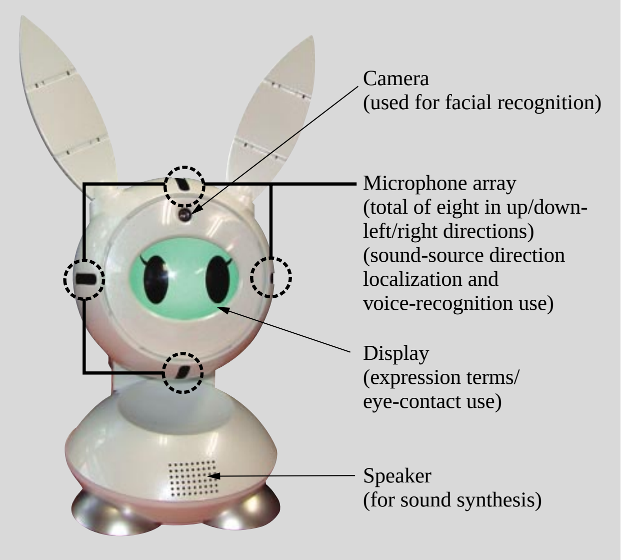
Camera
(used for facial recognition)
Microphone array
(total of eight in up/down-left/right directions)
(sound-source direction localization and voice-recognition use)
Display
(expression terms/ eye-contact use)
Speaker
(for sound synthesis)

Fig. 6—Structural Outline of Prototype Desktop Mascot for Intelligent User Interface.
The mascot is externally fitted with eight microphones for voice recognition, a camera for face-recognition, a speaker for synthesized-voice playback, and a display for facial expressions.

As regards the TV program recommender, for scheduled program broadcasts, regularly watched programs and user interests are evaluated, and recommendations are judged either good or bad.

Multimodal Dialog Platform

We have developed a "multimodal dialog platform" that acts as mediator between a user and various devices and databases. This platform uses XML (extensible markup language) to describe "scenarios" for carrying out user requests. Changing these scenarios as one likes makes it possible to execute a multitude of functions.

In the present study, the above-described technologies—namely, voice recognition, sound synthesis, face recognition, and TV program recommender—are integrated. And by creating scenarios specialized for operation of AV devices, an "intelligent user interface" has been created.

TEST PRODUCTION

A prototype version of the intelligent user interface incorporating the above-described interface technologies was manufactured in the form of a "desktop mascot" with the appearance of a rabbit (see

Fig. 6). The mascot is externally fitted with eight microphones for voice recognition and sound-source localization, a speaker for synthesized-voice playback, and a camera for face-recognition. Moreover, it incorporates a display, a movable head, and movable ears that can change shape. With these features, the mascot is able to render various expressions.

As well as continuing research using this prototype rabbit-type mascot, we are continuing investigations on incorporating all the functions described in this paper into the body of AV devices.

CONCLUSIONS

This paper overviewed the concept behind our "intelligent user interface"—which organically integrates media processing technologies aimed at realizing Hitachi's goal of a "user-friendly, high-definition-image lifestyle"—and describes the technologies incorporated in this interface. From now onwards, while increasing the number of degrees of freedom of the dialog interface, we are aiming for even greater variety of operation. To do so, at Hitachi, we will continue to further improve various elemental technologies, strive toward intelligent control of not only AV devices but all home information appliances by utilizing information retrieval from the Internet via home networks, and extend these functions into other areas.

REFERENCES

(1) M. Togami et al., "Adaptation Methodology for Minimum Variance Beam-Former Based on Frequency Segregation," Proc. of the 2005 Autumn Meeting of the Acoustical Society of Japan, 2-2-20, Acoustical Society of Japan (Sep. 2005) in Japanese.

(2) H. Kokubo et al., "Robust Speech Recognition for Car Environment Noise," *Electronics and Communications in Japan*, Part 3, Vol. 85, No. 11 (Nov. 2002).

(3) N. Nukaga et al.; "Unit Selection Using Pitch Synchronous Cross Correlation for Japanese Concatenative Speech Synthesis," 5th ISCA Speech Synthesis Workshop (Jan. 2004).

ABOUT THE AUTHORS

Hiroshi Shinjo

Joined Hitachi, Ltd. in 1990, and now works at the Intelligent Media Systems Research Department, the Central Research Laboratory. He is currently engaged in the research of image recognition systems. Mr. Shinjo is a member of The Institute of Electronics, Information and Communication Engineers (IEICE), and can be reached by e-mail at: shinjo@crl.hitachi.co.jp

Ui Yamaguchi

Joined Hitachi, Ltd. in 1999, and now works at the Intelligent Media Systems Research Department, the Central Research Laboratory. He is currently engaged in the development of robotics and human-machine interfaces. Mr. Yamaguchi is a member of The Japan Society for Precision Engineering (JSPE), and can be reached by e-mail at: ui-yama@crl.hitachi.co.jp

Akio Amano

Joined Hitachi, Ltd. in 1981, and now works at the Intelligent Media Systems Research Department, the Central Research Laboratory. He is currently engaged in the research and development of speech-recognition technology. Mr. Amano is a member of The Acoustical Society of Japan (ASJ), and can be reached by e-mail at: amano@crl.hitachi.co.jp

Konagi Uchibe

Joined Hitachi, Ltd. in 1992, and now works at the Intelligent Media Systems Research Department, the Central Research Laboratory. She is currently engaged in the development of recommender systems. Ms. Uchibe can be reached by e-mail at: konagi@crl.hitachi.co.jp

Atsushi Ishibashi

Joined Hitachi, Ltd. in 1983, and now works at the Home Solution Design Group, the Design Division. He is currently engaged in design of DVD cameras. Mr. Ishibashi can be reached by e-mail at: a-ishibashi@design.hitachi.co.jp

Hideki Kuwamoto

Joined Hitachi, Ltd. in 1987, and now works at the R&D Strategy Planning Department, the Business Planning Office, the Ubiquitous Platform Systems. He is currently engaged in the R&D planning of digital media products. Mr. Kuwamoto is a member of IEICE, and can be reached by e-mail at: hideki.kuwamoto.vk@hitachi.com