International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

F.746.9

(05/2019)

SERIES F: NON-TELEPHONE TELECOMMUNICATION SERVICES

Multimedia services

Requirements and architecture for indoor conversational robot systems

Recommendation ITU-T F.746.9

ITU-T

ITU-T F-SERIES RECOMMENDATIONS

NON-TELEPHONE TELECOMMUNICATION SERVICES

TELEGRAPH SERVICE	
Operating methods for the international public telegram service	F.1–F.19
The gentex network	F.20–F.29
Message switching	F.30–F.39
The international telemessage service	F.40–F.58
The international telex service	F.59–F.89
Statistics and publications on international telegraph services	F.90–F.99
Scheduled and leased communication services	F.100–F.104
Phototelegraph service	F.105–F.109
MOBILE SERVICE	
Mobile services and multidestination satellite services	F.110–F.159
TELEMATIC SERVICES	
Public facsimile service	F.160–F.199
Teletex service	F.200–F.299
Videotex service	F.300–F.349
General provisions for telematic services	F.350–F.399
MESSAGE HANDLING SERVICES	F.400–F.499
DIRECTORY SERVICES	F.500–F.549
DOCUMENT COMMUNICATION	
Document communication	F.550–F.579
Programming communication interfaces	F.580–F.599
DATA TRANSMISSION SERVICES	F.600–F.699
MULTIMEDIA SERVICES	**F.700–F.799**
ISDN SERVICES	F.800–F.849
UNIVERSAL PERSONAL TELECOMMUNICATION	F.850–F.899
ACCESSIBILITY AND HUMAN FACTORS	F.900–F.999

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T F.746.9

Requirements and architecture for indoor conversational robot systems

Summary

Recommendation ITU-T F.746.9 describes the requirements and architecture for an indoor conversational robot system. This Recommendation defines different functions to support indoor conversational robot systems. The scope of this Recommendation is focused on architecture, terminals, servers and interfaces between modules. It also describes different functions including a variable-distance speech preprocessing function, user identifying speech recognition function, multitask dialogue processing function and speech synthesis function. Additionally, it describes interfaces between modules for indoor conversational robot systems.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T F.746.9	2019-05-14	16	11.1002/1000/13916

Keywords

AI robot, dialogue processing, speech interface

[*] To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T F.746.9

Requirements and architecture for indoor conversational robot systems

1 Scope

This Recommendation provides an overview of the framework for indoor conversational robot systems. It describes the features, general requirements, architecture and functionality, which are the framework to support indoor conversational robot systems. The scope is focused on architecture, terminals, servers, interface and communication protocols between servers and clients in a high-level approach. This Recommendation will allow users of the indoor conversational robot system to be able to experience convenient information services in various areas with a user-friendly speech-dialogue interface.

2 References

None.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 co-reference resolution [b-ITU-T F.746.3]: A function that detects the preceding referents of the pronouns which replace the noun phrases of the input sentences.

3.1.2 named entity recognition [b-ITU-T F.746.3]: A function that recognizes named entities such as PLO which are people, organizations and places from the sentences. The PLO can be decomposed to more specific named entities depending on the applications.

3.1.3 natural language processing [b-ITU-T F.746.3]: A method that analyses text in natural languages through several processes such as part-of- speech recognition, syntactic analysis and semantic analysis.

3.1.4 part-of-speech recognition [b-ITU-T F.746.3]: A function that recognizes parts of speech (POS) in the sentences and assigns relevant POS tags considering contextual meaning of the target sentences.

3.1.5 semantic analysis [b-ITU-T F.746.3]: A function that recognizes the semantic relations among the words around predicates that exist in the same sentence. The semantic analysis function then generates a semantic predicate-argument structure (PAS).

3.1.6 speech [b-ITU-T H.703]: Speech is the vocalized form of human communication.

3.1.7 speech recognition [b-ITU-T H.703]: A kind of user interface for translation of spoken words into text.

3.1.8 syntactic analysis [b-ITU-T F.746.3]: A function that analyses sentence structures and generates dependency relations among words based on dependency grammars.

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 dialogue-based speech interface: An interface based on speech, especially dialogues between a user and a device or system.

3.2.2 dialogue act: The user's intention or purpose of the utterances in a dialogue, for example, a request for information, command for action or agreement.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
DA	Dialogue Act
DB	Database
DNN	Deep Neural Network
EER	Equal Error Rate
FAR	False Alarm Rate
FRR	False Rejection Rate
HCI	Human Computer Interaction
ICT	Information and Communication Technology
IPTV	Internet Protocol Television
IT	Information Technology
LLS	Language Learning System based on speech/NLP technology
LSTM	Long Short Term Memory
NE	Named Entity
NLP	Natural Language Processing
PC	Personal Computer
POS	Part Of Speech
RT	Reverberation Time
SMS	Short Message Service
SNR	Signal to Noise Ratio
SVM	Support Vector Machines
TTS	Text To Speech
VAD	Voice Activity Detection
VoD	Video on Demand

5 Conventions

The following conventions are used in this Recommendation:

– The keywords "is required to" indicate a requirement that must be strictly followed and from which no deviation is permitted, if conformance to this Recommendation is to be claimed.

– The keywords "is prohibited from" indicate a requirement that must be strictly prohibited, if conformance to this Recommendation is to be claimed.

– The keywords "is recommended" indicate a requirement that is recommended but which is not absolutely required. Thus, this requirement need not be present to claim conformance.

– The keywords "is not recommended" indicate a requirement that is not recommended but which is not specifically prohibited. Thus, conformance with this Recommendation can still be claimed even if this requirement is present.

– The keywords "can optionally" indicate an optional requirement that is permissible, without implying any sense of being recommended. This term is not intended to imply that the vendor's implementation must provide the option and the feature can be optionally enabled by the network operator/service provider. Rather, it means the vendor may optionally provide the feature and still claim conformance with this Recommendation.

6 Introduction

These days the application of artificial intelligence (AI) technology has expanded further into areas such as robotics. Indoor conversational robot systems are one of the AI applications equipped with speech interfaces that enable users to have a dialogue with robots. The key technologies in such a multimedia system are speech recognition technology for variable distance speech and multitask dialogue processing technology which analyses the resulting natural language text from the speech recognition, understands the meaning of the spoken dialogues and generates the best relevant responses. In this way, the intelligent robot can perform the service for the human user while having a conversation with him/her.

The indoor conversational robot can be applied to various areas. It can serve people as a kiosk robot to guide customers or to introduce different services in the private and public sectors. It can also serve as a conversational helper for everyday life for individuals living alone, e.g. an elderly person. It can provide people with recipes through question and answer speech dialogues. It can be used as a personal assistant robot to manage a schedule. The last use-case example is a robot guide which provides dialogue-based information service on domestic multimedia content such as that for Internet Protocol TV (IPTV) or video on demand (VoD).

In this document, we present the requirements, architecture and functions for indoor conversational robot systems.

6.1 Basic concept of indoor conversational robot

An indoor conversational robot is a type of intelligent agent that has the capability to have a dialogue with people in a room. The purpose of the dialogue between a robot and people is for the robot to provide various services to people. To provide such intelligent services the core technologies include variable-distance speech recognition and multitask dialogue processing technology. The robot should first be able to recognize speech spoken by a person in a noisy room who may walk around. It means the distance of the speech recognition target is not determined but variable. After the speech recognition is successful, the robot then should be able to understand the result of the speech recognition that is in text form. It should process the text that is in a natural language to find the meaning of the speech in terms of different task dimensions. It also performs the dialogue processing by dialogue management and dialogue status tracking techniques to find out the speaker's intention and purpose of the dialogue while adapting to the speaker's conversational style. In the long run, the robot, or the intelligent agent achieves his goal by providing the appropriate service after figuring out the conversation.

7 Requirements of indoor conversational robot systems

The general requirements for an indoor conversational robot system (ICR) are described as follows.

– An ICR is required to provide a speech recognition function for spontaneous speech.

– An ICR is required to be able to recognize the predefined activation command spoken by the user.

- An ICR is recommended to provide a keyword registration function for the user.
- An ICR is recommended to provide a dialogue processing function for spontaneous utterances.
- An ICR is recommended to be able to understand the user's intention by analysing the speaker's speech acts.
- An ICR is recommended to provide a conversation tracking function to understand the interactive dialogues between the user and the system.
- An ICR is recommended to provide a dialogue generation function to perform the interactive dialogues between the user and the system.
- An ICR is recommended to be able to identify pre-validated users.
- An ICR is recommended to be able to recognize the location of the user and change its position towards the user.
- An ICR is recommended to be able to identify pre-validated users, recognize the location of the users and the predefined activation command at the same time.
- An ICR is recommended to recognize speech in the noisy environment with reverberation.
- An ICR is recommended to process dialogues demanding multiple tasks.
- An ICR is recommended to be able to adapt to the user through conversational experiences through incremental learning.

8 Functional components and interfaces for indoor conversational robot systems

The functional components of indoor conversational robot systems, presented in Figure 1, are composed of the following modules:

- variable-distance speech preprocessing module;
- user identifying speech recognition module;
- multitask dialogue processing module;
- speech synthesis module (TTS: text to speech).

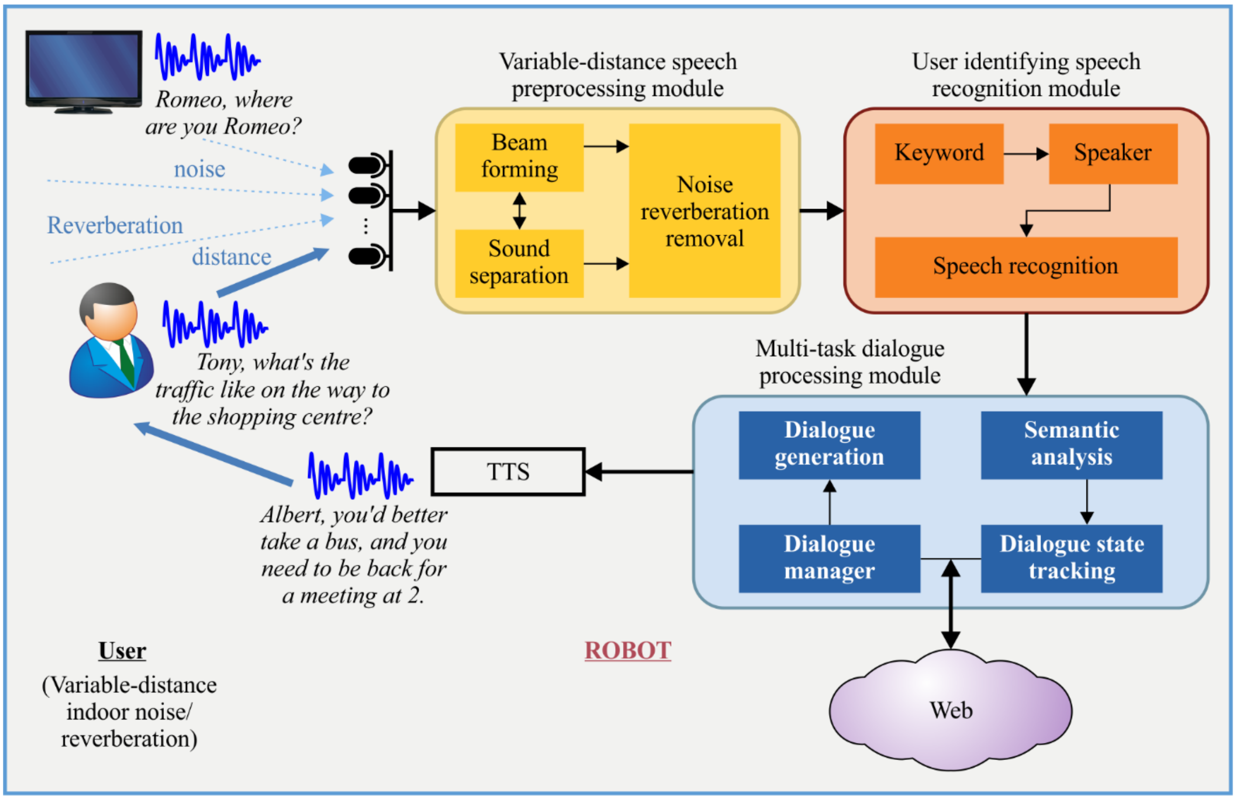


Figure 1 – Indoor conversational robot system

In the variable-distance speech preprocessing module, various reverberations of the indoor environments and noise inflicted variable-distance speech is cleaned and the reliable speech recognition is made possible.

In the user identifying speech recognition module, the pre-validated user calls the robot with predefined words which is waiting for a command. The robot is then activated, recognizes the specific user and starts a dialogue with the user.

In the multitask dialogue processing module, the robot understands the information in the natural language text which is the speech recognition result. The robot has the conversation with the user by adaptive dialogue management, generates responses and expands or changes the dialogue topic appropriately depending on the different tasks.

In the speech synthesis module, the result of the dialogue generation turns into the sound of the robot. That is the well-known text-to-speech module in the general speech interface.

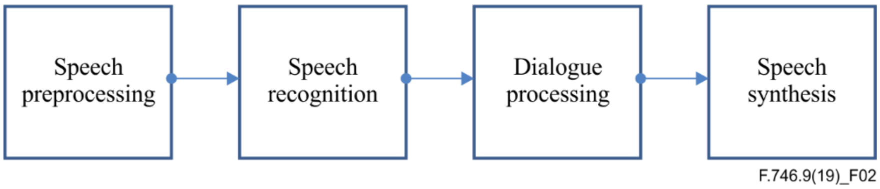


Figure 2 – Functional architecture of indoor conversational robot system

8.1 Variable-distance speech preprocessing module

In the variable-distance speech preprocessing module, various reverberations of the indoor environments and noise inflicted variable-distance speech is cleaned and the reliable speech recognition is made possible. The following functions are included in the variable-distance speech preprocessing module as essential functions:

– Sound localization function that is robust for reverberation – This function detects the direction of the sound source with noise and reverberation;

– Noise reverberation-removing function – This function removes or minimizes noise and reverberations from the sound source using methods such as the beamforming algorithm that keeps the direction information and minimizes noise from the sound. Reverberation suppression is performed which is independent of the array structure of the microphone.

– Sound separation function – This function separates voice features from multichannel input sound signal information of frequency, time and channel to remove background noise. The resulting voice signal proceeds to a speech recognizer as the input signal.

– Sound feature compensation function – This function changes from distorted voice feature to clean voice feature using inference by deep neural network (DNN) sound modelling for better speech recognition.

For reliable variable-distance speech recognition, the database (DB) information on the items in the following table is required.

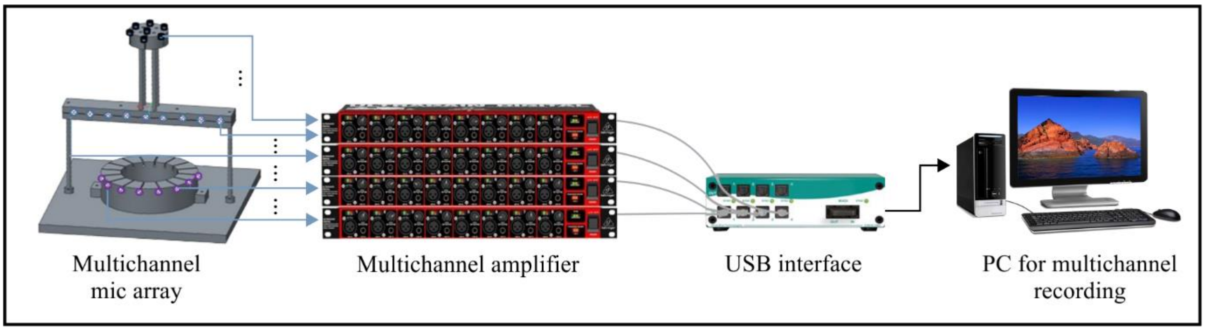
Table 1 – Database information for variable-distance speech recognition for indoor conversational robot system

DB information type	Information subtype	Example
Language information		
	Language type	English, French, Korean
Speaker information		
	Gender	Male, Female
	Age	17, 25, 60th, 30th
	Origin	Macao, Italy, Southern France
Room acoustic information		
	Surrounding noise type	Point noise source/ambient noise
	Noise type	Music, babble
	Speech-to-nose ratio	100 vs. 1
Microphone information		
	Model name	MT600
	Type of microphone	Dynamic, condenser, mems
	Channel type	Single/multiple
	Number of microphones	
	Microphone array configuration type	linear, circular
Speech information		
	Location of the speaker	Distance and angle from microphone
	Speech contents	Words, linked words, sentences
	Speech style	Reading, spontaneous
File information		
	File type	Raw, wav
	Sampling rate	8 kHz, 16 kHz, 44.1 kHz, 48 kHz
	File name structure	Speech1-raw-16 kHz

8.2 User identifying speech recognition module

The user identifying speech recognition module performs the following processes: the pre-validated user calls the robot with predefined words which is waiting for a command. The robot is then activated, recognizes the specific user, and starts a dialogue with the user. The function of the speech recognition module is to generate output text from the user's speech signals. The following functions are used for the user identifying speech recognition module.

– Speech DB construction – This function, for the indoor home environment, variable distance multichannel speech DB collection, is needed using microphone arrays for the speech recognition module. The collected speech utterances is arranged to include silent sections for end-point detection.



F.746.9(19)_F03

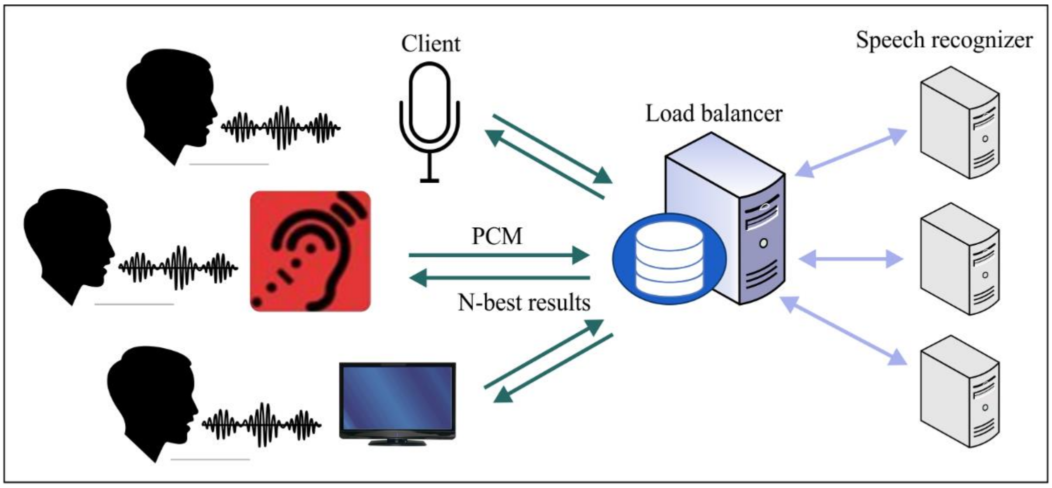
Figure 3 – Example of multichannel microphone array speech DB collection system

– End-point detection function – This function detects the voice signal from the input signal. This function is used as a preprocess for speech recognition, speaker verification or speaker identification. It consists of voice activity detection (VAD) stage and smoothing stage. Voice activity detection (VAD) is to classify the input signal into voice and non-voice using DNN. The result of VAD is run through a smoothing process by a hangover technique and returns the end-point of each sentence.

– Keyword extraction function – This function, which utilizes a DNN sound model, is robust against background noise. Keywords are registered by the user for activating robots and used as training data for keyword extraction and speech recognition.

– Speaker identification function – This function, which utilizes DNN feature analysis, is robust against background noise. Speaker identification is a type of classification task to detect a speaker of the utterance from a certain speaker group. Recent AI technologies such as DNN improves the performance of classification tasks.

– Automatic speech recognition function (ASR) – This function generates output text from the user's speech signals. The ASR function should be scalable to multi-robot environments as the example architecture in Figure 4 shows. Training data for ASR is composed of that recorded in similar environments with noise and variable distance.



F.746.9(19)_F04

Figure 4 – Speech Interface for multiple conversational robots

The user identifying speech recognition module employs the deep neural network (DNN) model in most functions in the module. In general, a deep neural network (DNN) model shows good performance in the generation or classification task. However, DNN fundamentally has non-linear characteristics, and therefore it is difficult to interpret how a result from inputs given to a black box model has been derived. To explain the hidden process, a linear approximation of a DNN model which is a non-linear function is performed with the following operations consecutively: an operation of expanding an input to a neuron of the DNN into a polynomial; a second operation of approximating the neuron of the DNN with a Taylor series in parallel with the polynomial expansion of the input; and a third operation of classifying the polynomially-expanded input and the Taylor-series approximated neuron as a polynomial of input signals and a polynomial of weights.

8.3 Multitask dialogue processing module

The multitask dialogue processing module performs the following functions: the robot understands the information in the natural language text which is the speech recognition result. The robot has the conversation with the user by adaptive dialogue management, generates responses and expands or changes the dialogue topic appropriately depending on the different tasks. The following functions are included in the multitask dialogue processing module.

- Semantic analysis of dialogues – Semantic analysis is the essential function for understanding the user's intention of the utterance. It should be capable of multitask semantic tagging which is used for the analysis. Speech act and dialogue act analysis models are employed for the dialogue understanding function. A slot filling model for dialogue understanding is constructed based on DNN using a large corpus of user dialogues. Dialogue topic expansion and modification to improve the dialogue model are also included in the semantic analyser function.
- Dialogue state tracking function – This function detects the state of each turn of the user/robot dialogue such as "question", "answer" or "making commands" which are robust for dialogue topic expansion and modification. DNN-based dialogue state tracking technology is used for dialogue state recognition. In Figure 5 an example dialogue state tracker is presented which uses long short-term memory (LSTM) networks, a recently developed neural network.

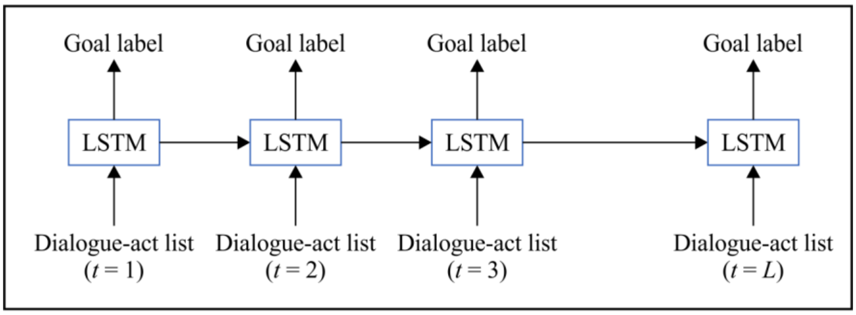

Goal label
Goal label
Goal label
Goal label
LSTM
LSTM
LSTM
LSTM
Dialogue-act list (t = 1)
Dialogue-act list (t = 2)
Dialogue-act list (t = 3)
Dialogue-act list (t = L)
F.746.9(19)_F05

Figure 5 – Example of dialogue state tracker

- Dialogue management function – This function manages the dialogue between the user and the robot based on the intention of the user. The dialogue management function should be capable of expanding and modifying dialogue topics. Various reinforcement learning methods based on DNN are employed for better performance of dialogue management.
- Dialogue generation function – This function produces a robot utterance with a correct dialogue act from the input dialogue act of the human user. Based on a statistical model the dialogue generation function is task adaptable for multiple dialogue topics.

8.4 Speech synthesis module

The speech synthesis module performs the following functions: the result of the dialogue generation turns into the sound of the robot. This is the well-known text-to-speech module in the general speech interface. The speech synthesis module generates speech signals from the text input by text normalization and morphological analysis, intonation target (i.e., pitch contour) generation and speech signal generation using language analysis and intonation prediction information for the input sentences.

9 Application programme interfaces (API) for indoor conversational robot systems

In this clause, APIs for different modules of a robot system are defined: communication between client and server, preprocessing module, speech recognition module and dialogue processing module.

9.1 APIs to communicate between client and server

The APIs to communicate between the client and server consists of methods to make a new connection to the server, to recognize recorded multichannel signals, to get an ASR or dialogue result for a given input signal and to get a response from the robot to the current user input in wave format.

Table 2 – APIs for communication between server and client

Constructor	
Constructor and description	
ConversationalRobot()	
Default constructor	
Methods	
Modifier and type	Method and description
int	openConnection(String id, String ipAddress, Integer port)
	Make a new connection to ASR/dialogue server id: Unique identifier of client ipAddress: ip address of ASR/Dialogue server port: TCP port number to connect server

Table 2 – APIs for communication between server and client

	Returns 0 on success or ERROR code otherwise
void	closeConnection()
int	setParameters(String key, String value) Set control parameters of either client or server including number of channels, kind of input signal, domain of conversation etc key: parameter name value: parameter value Returns 0 on success or ERROR code otherwise
String	getParameters(String key) Get control parameters of either client or server key: parameter name Returns value corresponding to key
int	Recognize(Array<Array<sig_type>> sig, int dim, int len)
	Recognize recorded multichannel signal dim: number of microphone channels len: length of signal Returns status of recognizer
String	getResult(int kind) Get ASR or dialogue result for given input signal so far kind: kind of result: ASR or Dialogue Returns result of ASR, dialogue, etc in text format
Array<sig_type>	getResponse() Get response of robot to current user input in wave format, if any Returns waveform data which can be played in
String	getStatus(int kind) Get current status of ASR or dialogue system kind: kind of status Returns the current status of engine in text format

9.2 APIs for preprocessing module

The APIs for the preprocessing module creates resources for the preprocessing and processing of a block of multichannel audio signals for the use of speech recognition in the next step.

Table 3 – APIs for preprocessing module

Constructor	
Constructor and description	
CPreproc()	
Default constructor	
Methods	
Modifier and type	Method and description
int	createPreproc(int ch, int fs) Create resources for preprocessing ch: number of channels fs: sampling frequency of input signal Returns 0 on success, error code otherwise

Table 3 – APIs for preprocessing module

void	freePreproc() Free allocated resources
int	processBlock(Array<Array<sig_type>> insig, int len, Array<sig_type>& outsig) Preprocess a block of muti-channel audio signals insig: multichannel speech signals len: length of input signals outsig: output single channel signal Returns length of output signals
int	setParameters(String key, String value) Set control parameters of preprocessing module key: parameter name value: parameter value Returns 0 on success or error code otherwise
String	getParameters(String key) Get control parameters of preprocessing module key: parameter name Returns value corresponding to the key

9.3 APIs for automatic speech recognition module

This module includes key-word spotting, speaker identification, language identification and speech recognition. It returns result of the ASR module in a text format. The result can have more than one hypothesis.

Table 4 – APIs for ASR module

Constructor	
Constructor and description	
CASR()	
Default constructor	
Methods	
Modifier and type	Method and description
int	createASR(int ch, int fs) Create resources for ASR ch: number of channels fs: sampling frequency of input signal Returns 0 on success, error code otherwise
void	freePreproc() Free allocated resources
int	processBlock(Array<Array<sig_type>> insig, int len) Process a block of multichannel audio signals for key-wordspotting, speaker identification, language identification, speech recognition etc insig: multichannel speech signals len: length of input signals

Table 4 – APIs for ASR module

	Returns status of engine
int	setParameters(String key, String value) Set control parameters of ASR module key: parameter name value: parameter value Returns 0 on success or error code otherwise
String	getParamters(String key) Get control parameters of ASR module key: parameter name Returns value corresponding to key
Array<String>	getResult(int kind) Get result of ASR module so far kind: a kind of result requested Returns result of ASR module in text format. The result can have more than one hypothesis

9.4 APIs for dialogue module

The APIs for the dialogue module create resources for dialogue processing and processing text for the multiple candidate results. It returns the result of the dialogue module in a text format. The result can have more than one hypothesis as with the speech recognition module.

Table 5 – APIs for dialogue module

Constructor	
Constructor and description	
CDialog()	
Default constructor	
Methods	
Modifier and type	Method and description
int	createDialog(int lang) Create resources for dialogue lang: language id Returns 0 on success, error code otherwise
void	freeDialog() Free allocated resources
int	processDialog(Array<String> nbest) Process a text of nbest hypothesis nbest: contents of user utterance Returns status of engine
int	set Parameters(String key, String value) Set control parameters of dialogue module key: parameter name value: parameter value Returns 0 on success or error code otherwise

Table 5 – APIs for dialogue module

String	getParameters(String key) Get control parameters of dialogue module key: parameter name Returns value corresponding to key
Array<String>	getResult(int kind) Get result of dialogue module so far kind: a kind of result requested Returns result of dialogue module in text format. The result can have more than one hypothesis

Appendix I

Evaluation method for measuring performance of an indoor conversational robot system in variable distant noisy environments

(This appendix does not form an integral part of this Recommendation.)

In this appendix the evaluation method for evaluating an indoor conversational robot system is introduced. The evaluation method consists of the three methods given in clauses I.1, I.2 and I.3, each of which evaluates different function.

I.1 Evaluation of speech recognition in variable distant noisy environments

The speech recognition performance is evaluated in a home-like space with various noise sources where the user speaks given sentences. Word error rate (WER) is the evaluation metric for the speech recognition performance. WER is measured by comparing the sentences spoken by the user with those recognized by the speech recognition system. The WER calculation formula is as follows [b-Klakow].

$$\mathrm{WER} = (S + D + I)/N$$

S = number of substituted words

D = number of deleted words

I = number of inserted words

N = total number of words in reference text

Noise sources that are considered are home environments such as TV, audio sounds, voice babble and noise from home appliances. Noise volume is decided based on the signal to noise ratio (SNR) by comparing it with the user's voice volume. The performance of the speech recognition function is measured after noise sources are produced according to three conditions as SNR 15 dB, 10 dB, 5 dB. The size of the space, reverberation feature, location of the robot, location of the sound source (relative distance and direction of the sound source and robot) and location of the noise source (relative distance and direction of the noise source and robot) are considered for the evaluation of the speech recognition function as presented in Figure A.1 as an example. Reverberation time (RT) is a measure of the time required for the sound to "fade away" in an enclosed area after the source of the sound has stopped. The RT60 metric is used for reverberation feature of the space which expresses the degree of reverberation. For the RT60 metric, 15~20 ms is used to mimic general home environments. RT60 is defined as the time it takes for the sound pressure level to reduce by 60 dB, measured after the generated test signal is abruptly ended.

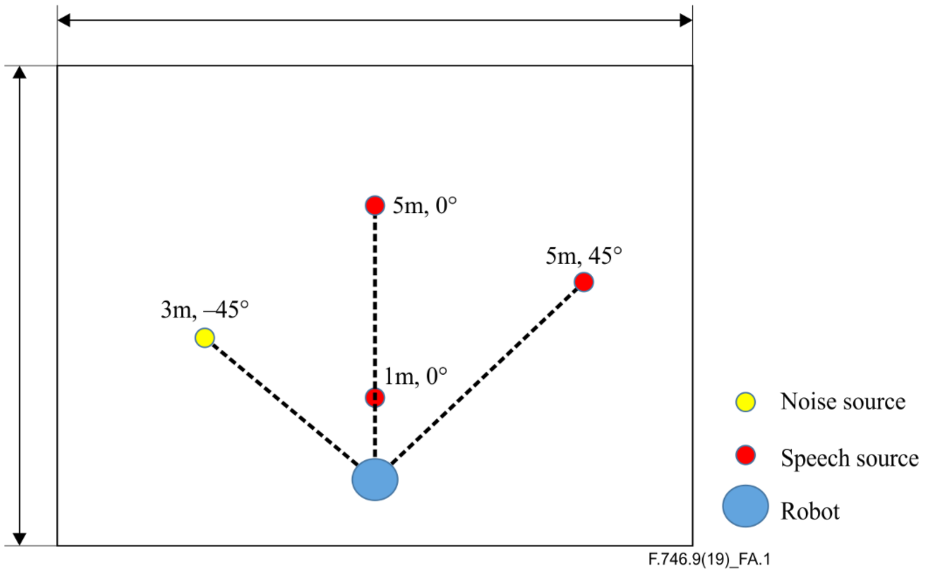


Figure A.1 – Example configuration of microphones, speech and noise source

I.2 Evaluation of call success in noisy environments

Call success rate measures the performance of a robot who recognizes the voice call successfully and gets ready for the response. False rejection rate (FRR), false alarm rate (FAR), and equal error rate (EER) are used as call success rate metrics. FRR is defined as the rate that a robot fails to recognize the predefined word or phrases a user speaks. Therefore, the call success rate is [1-FRR].

FRR = (number of missed calls)/(total calls tried)

False recognition voice calls which are not correct call words or phrases are defined by FAR.

FAR = (number of false recognized calls)/(total calls tried)

Recognized calls are accepted or rejected based on the threshold value that was predefined by the statistical model. The performance of FRR and FAR is decided depending on the threshold tuning state. In general, FAR increases if FRR is improved. Considering this general case, the performance can be measured using EER where FRR and FAR are set to the same value. The selection of FRR, FAR, EER can be decided according to the application purpose of the robot.

I.3 Evaluation of conversations of specific purposes in indoor noisy environments

Conversations are evaluated by the conversation success rate which measures the degree of achievement of the defined missions that a robot and a human user try to complete through speech conversations. The conversation success rate is measured by the number of successful missions among the total missions that were achieved within the defined number of turns of the conversation.

An example would be the case where the number of conversation turns are limited to within 5 turns. The mission of "Make the dental reservation with a dentist at 3 p.m." is regarded as successful if the robot finishes the mission within 5 turns.

Bibliography

[b-ITU-T F.746.3] Recommendation ITU-T F.746.3 (2015), *Intelligent question answering service framework*.

[b-ITU-T H.703] Recommendation ITU-T H.703 (2016), *Enhanced user interface framework for IPTV terminal devices*.

[b-Klakow] Klakow, D. & Peters, J., 2002, *Testing the correlation of word error rate and perplexity*. Speech Communication. Vol.38, No.1; pp. 19–28.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	**Non-telephone telecommunication services**
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2019