International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

H.862.5

(06/2021)

SERIES H: AUDIOVISUAL AND MULTIMEDIA SYSTEMS

E-health multimedia systems, services and applications – Multimedia e-health data exchange services

Emotion enabled multimodal user interface based on artificial neural networks

Recommendation ITU-T H.862.5

ITU-T H-SERIES RECOMMENDATIONS

AUDIOVISUAL AND MULTIMEDIA SYSTEMS

CHARACTERISTICS OF VISUAL TELEPHONE SYSTEMS	H.100–H.199
INFRASTRUCTURE OF AUDIOVISUAL SERVICES	
General	H.200–H.219
Transmission multiplexing and synchronization	H.220–H.229
Systems aspects	H.230–H.239
Communication procedures	H.240–H.259
Coding of moving video	H.260–H.279
Related systems aspects	H.280–H.299
Systems and terminal equipment for audiovisual services	H.300–H.349
Directory services architecture for audiovisual and multimedia services	H.350–H.359
Quality of service architecture for audiovisual and multimedia services	H.360–H.369
Telepresence, immersive environments, virtual and extended reality	H.420–H.439
Supplementary services for multimedia	H.450–H.499
MOBILITY AND COLLABORATION PROCEDURES	
Overview of Mobility and Collaboration, definitions, protocols and procedures	H.500–H.509
Mobility for H-Series multimedia systems and services	H.510–H.519
Mobile multimedia collaboration applications and services	H.520–H.529
Security for mobile multimedia systems and services	H.530–H.539
Security for mobile multimedia collaboration applications and services	H.540–H.549
VEHICULAR GATEWAYS AND INTELLIGENT TRANSPORTATION SYSTEMS (ITS)	
Architecture for vehicular gateways	H.550–H.559
Vehicular gateway interfaces	H.560–H.569
BROADBAND, TRIPLE-PLAY AND ADVANCED MULTIMEDIA SERVICES	
Broadband multimedia services over VDSL	H.610–H.619
Advanced multimedia services and applications	H.620–H.629
Content delivery and ubiquitous sensor network applications	H.640–H.649
IPTV MULTIMEDIA SERVICES AND APPLICATIONS FOR IPTV	
General aspects	H.700–H.719
IPTV terminal devices	H.720–H.729
IPTV middleware	H.730–H.739
IPTV application event handling	H.740–H.749
IPTV metadata	H.750–H.759
IPTV multimedia application frameworks	H.760–H.769
IPTV service discovery up to consumption	H.770–H.779
Digital Signage	H.780–H.789
E-HEALTH MULTIMEDIA SYSTEMS, SERVICES AND APPLICATIONS	
Personal health systems	H.810–H.819
Interoperability compliance testing of personal health systems (HRN, PAN, LAN, TAN and WAN)	H.820–H.859
Multimedia e-health data exchange services	**H.860–H.869**
Safe listening	H.870–H.879

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

Recommendation ITU-T H.862.5

Emotion enabled multimodal user interface based on artificial neural networks

Summary

Recommendation ITU-T H.862.5 provides functional entities and architecture for emotion enabled multimodal user interface based on artificial neural network.

As emotion technology continues to make big improvements in human-computer interaction (HCI) areas, many companies and researchers have been studying emotion technology. Various applications using multimodality and emotion analysis are also introduced these days with artificial intelligence technology. However, many of the current systems do not yet infer human emotion properly because some systems are either too dependent on certain sources, or too weak for real circumstances.

Therefore, the proposed system architecture is for multimodal user interface (UI) based on emotion analysis with some properties and illustrations, and data with an artificial neural network. The multimedia data for the input is composed of text, speech, and image. For the unimodal emotion analysis, the data is pre-processed in the corresponding module. For example, the text data is pre-processed by data augmentation, person attributes recognition, topic cluster recognition, document summarization, named entity recognition, sentence splitter, keyword cluster, and sentence to graph functions.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T H.862.5	2021-06-13	16	11.1002/1000/14690

Keywords

Emotion expansion, multimedia knowledge database, multimodal emotion analysis, unimodal emotion analysis.

* 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/software copyrights, 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 appropriate ITU-T databases available via the ITU-T website at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Recommendation ITU-T H.862.5

Emotion enabled multimodal user interface based on artificial neural networks

1 Scope

This Recommendation describes the functional entities and architecture for emotion enabled multimodal user interface based on artificial neural networks.

In particular, the scope of this Recommendation includes:

– Architectural framework;

– Functional entities;

– Interfaces;

– Application to multimodal emotion analysis.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

None.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

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

3.1.2 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.

3.1.3 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).

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

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

3.1.6 knowledge base [b-ITU-T F.746.7]: A collection of knowledge resources that consist of structured and unstructured data. The knowledge base is used to provide information to the various applications that are related to information provisioning such as question answering (QA) systems and search systems.

3.1.7 question answering [b-ITU-T F.746.7]: A system that provides answers in a natural language to questions which are in the natural language form by analysing the questions and all the knowledge resources that are available to the system.

3.2 Terms defined in this Recommendation

None.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

CNN Convolutional Neural Networks

LSTM Long Short-Term Memory

NE Named Entity

NN Neural Network

POS Part of Speech

QA Question Answering

STT Speech-to-Text

UI User Interface

5 Conventions

None.

6 Functional architecture

6.1 Architectural framework

Figure 1 shows the proposed system architecture for multimodal user interface (UI) based on emotion analysis with some properties and illustrations. The multimedia data is composed of text, speech, and image. For the unimodal emotion analysis, the data is pre-processed in each analysis module. In other words, after the pre-processing stage, three unimodal networks analyse each data. The multimodal network then integrates the state vectors that denote recognized emotion from each unimodal process. The emotion obtained by the multimodal network expands to complex emotion through a complex emotion generation module.

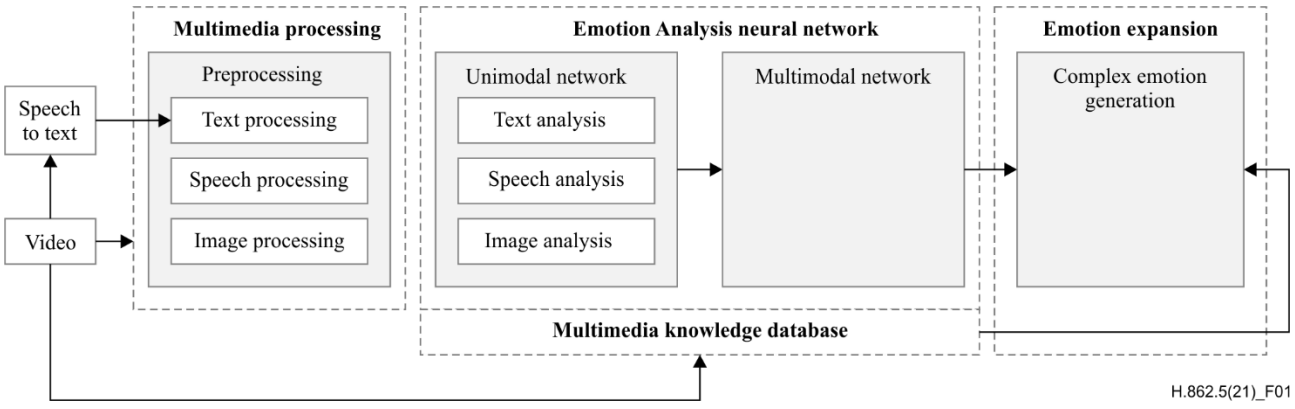

Figure 1 – Emotion enabled multimodal UI architecture

The proposed UI architecture presents the emotion analysis processes for how to pre-process emotional sources and integrate unimodal networks. The system composing proper methods is a key factor for multimodal emotion analysis.

7 Functional entities

7.1 Multimedia processing – pre-processing module

In the pre-processing module low-level handcrafted features are produced from text input, audio input and image input, respectively. That is, the data pre-processing module performs the embedding function of transforming the text data, audio data and image data into vector types that can be used for machine learning. The text input is produced from the audio input by STT (speech-to-text) or speech recognition module.

7.1.1 Text processing

As shown in Figure 2, the text data is pre-processed by a natural language processing method that includes text data augmentation, person attributes recognition for text, topic cluster recognition, document summarization, named entity (NE) recognition, sentence splitter, and keyword cluster sub-functions.

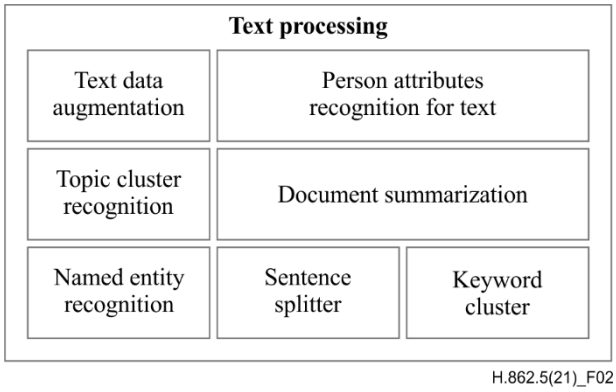

Figure 2 – Text processing module

7.1.1.1 Text data augmentation

The role of the text data augmentation function is to increase the training text data to improve the performance of the emotion recognition network. It takes the original text data as input and produces/returns the new training data using various augmentation methods such as the thesaurus, word embedding, and back translation. The natural language sentences are expanded for the trainer through an expansion tool that takes the input sentences and analyses them for part of speech (POS) tagging. The primary parts of speech in the sentences are expanded to the secondary POS through a search for similar words in the database or in the synonym dictionary. The input sentences are then expanded by replacing them with those of the extracted secondary POS. Finally, the newly generated input sentences used for training will then be verified for grammatical correctness.

– Input: text data
– Output: text data

7.1.1.2 Person attribute recognition for text

The person attribute recognition for text function finds speakers in the text, analyses the characteristics of the speakers and recognizes attributes such as gender and age.

– Input: text data
– Output: attributes of the speaker

7.1.1.3 Topic cluster recognition

Topic recognition function recognizes topics of the text and calculates the importance of each topic by analysing the statistics of the words mathematically.

– Input: text data

– Output: topic cluster of the text data and their importance measures

7.1.1.4 Document summarization

Document summarization function divides the original text into multiple blocks of phrases and extracts the summarized sentence which represents each block of phrases.

– Input: text data

– Output: summarized sentences

7.1.1.5 Named entity recognition

The named entity recognition function extracts nouns in the text data which is provided by a part-of-speech analyser. It then recognizes the named entities for the extracted nouns.

– Input: text data

– Output: nouns in the text and corresponding named entities

7.1.1.6 Sentence splitter

Sentence splitter separates each sentence from the input text data that may consist of multiple sentences.

– Input: text data

– Output: sentences in the text data

7.1.1.7 Keyword cluster

Keyword cluster function makes groups for the keywords which are extracted from the text data based on relation description features among the keywords such as a similarity feature.

– Input: text data

– Output: keyword cluster

7.1.2 Speech processing

As shown in Figure 3, speech data is pre-processed by speech data augmentation, person attributes recognition for speech, noise reduction, voice separation, and voice activity detection sub-functions.

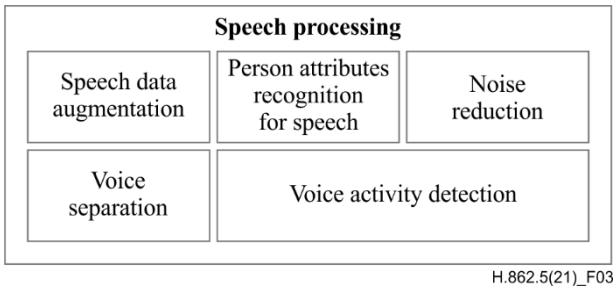

Figure 3 – Speech processing module

7.1.2.1 Speech data augmentation

Speech data augmentation function performs the role of increasing the training data for the performance enhancement of the emotion recognition network. It takes the original speech data and produces new training data using various augmentation methods such as time warping and frequency masking.

– Input: speech data

– Output: speech data

7.1.2.2 Person attributes recognition for speech

Person attributes recognition for speech function finds the speaker in the speech data, analyses his features and recognizes features such as gender and age.

– Input: speech data

– Output: attributes of the speaker

7.1.2.3 Noise reduction

Noise reduction function removes from the speech data various noises such as background sound, white noise, and third person utterances.

– Input: speech data

– Output: speech data

7.1.2.4 Voice separation

Voice separation function separates the background sound and voice from the speech data.

– Input: speech data

– Output: voice and background sound

7.1.2.5 Voice activity detection

Voice activity detection function removes the non-linguistic meaningless section in the speech data.

– Input: speech data

– Output: voice without non-linguistic meaningless section

7.1.3 Image processing

Figure 4 shows that the image data is pre-processed by image data augmentation, person attributes recognition for image, noise reduction, object detection, face detection, and gesture recognition sub-functions.

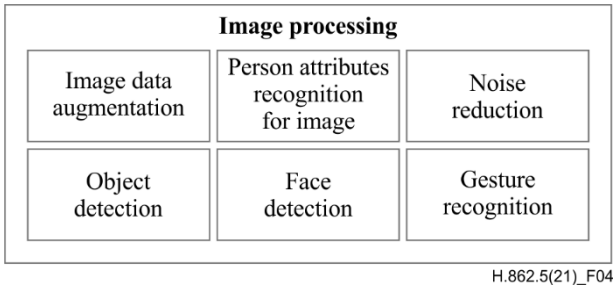

Figure 4 – Image processing module

7.1.3.1 Image data augmentation

Image data augmentation function performs the role of increasing the training data for the performance enhancement of the emotion recognition network. It takes the original image data and produces new training data using various augmentation methods such as cropping, translation, and flipping.

– Input: image data

– Output: image data

7.1.3.2 Person attributes recognition for image

Person attributes recognition for image function finds the person in the image data, analyses individual features, and recognizes features like gender, age, and clothes of the person.

– Input: image data

– Output: location information and other attributes of a person

7.1.3.3 Noise reduction

Noise reduction function does the role of removing noise from the image data. This function also generates training data after the sensor noises and illuminance disparities are removed.

– Input: image data

– Output: image data

7.1.3.4 Object detection

Object detection function classifies the backgrounds in the image data that can be used for emotion recognition tasks. It produces the coordinate values of the relevant pixels and classification results.

– Input: image data

– Output: location information of the background image and classification results

7.1.3.5 Face detection

Face detection function finds a persons' face in the image data and produces the pixel coordinate values.

– Input: image data

– Output: location information of the found faces

7.1.3.6 Gesture recognition

Gesture recognition function recognizes the ongoing gestures from a single image or consecutive images taken at a certain time interval. It returns the recognition results as the output.

– Input: single or consecutively taken image data

– Output: gesture classification results

7.2 Emotion analysis neural network – unimodal network

After the pre-processing stage, three unimodal networks analyse each data as shown in Figure 5. And then, the multimodal network integrates state vectors from each unimodal process.

The text emotion inference engine recognizes the inner emotion embedded in the text input from the low-level features using the neural network (NN) which consists of an input layer, at least two hidden layers and a softmax layer. The output of the softmax layer is the emotion vector. The emotion vector is composed of probabilities of seven types of basic emotions such as happiness, sadness, fear, disgust, anger, surprise, and neutral [b-Ekman]. One of the hidden layers produces, as the output, high-level features from the low-level features by dimensionality reduction. The text emotion inference engine produces one hidden layer output of the neural network as a state vector which will be used as the input to the multimodal network.

Like the text emotion inference engine, the speech emotion inference engine and the image emotion inference engine infer the innate emotion from the low-level features in the audio and image input, using the neural network and produces the state vector to be input to the multimodal network. The text emotion inference engine and the speech emotion inference engine can use the long short-term memory (LSTM) network. The image emotion inference engine can use the convolutional neural network (CNN). Each unimodal module has a trainer module and trains the related neural network of the emotion inference engine at the training stage.

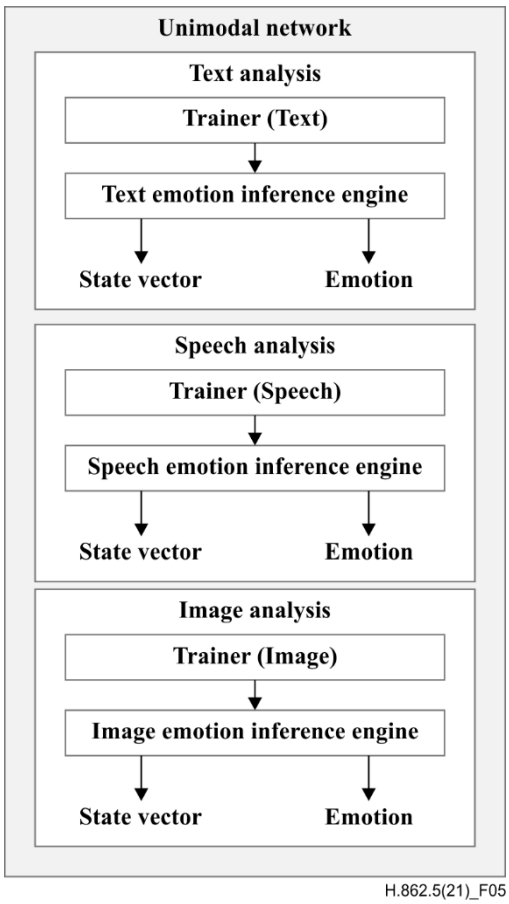

Figure 5 – Unimodal network module

7.2.1 Text analysis

7.2.1.1 Trainer for the text analysis

The trainer for the text analysis trains the neural networks using the text data information produced in the pre-processing stage in clause 7.1.1.

– Input: pre-processed text data

– Output: text emotion inference engine

7.2.1.2 Text emotion inference engine

Text emotion inference engine analyses the input text data information and produces emotion and state vector result value using the neural network from the trainer.

– Input: pre-processed text data

– Output: state vector, emotion

7.2.2 Speech analysis

7.2.2.1 Trainer for speech analysis

The trainer for the speech analysis trains the neural network using the speech data information produced in the pre-processing stage described in clause 7.1.2.

– Input: pre-processed speech data

– Output: speech emotion inference engine

7.2.2.2 Speech emotion inference engine

Speech emotion inference engine analyses the input speech data information and produces emotion and state vector result value using the neural network from the trainer.

– Input: pre-processed speech data

– Output: state vector, emotion

7.2.3 Image analysis

7.2.3.1 Trainer for image analysis

The trainer for image analysis trains the neural network with the image data information generated in the image processing stage described in clause 7.1.3.

– Input: pre-processed image data

– Output: text emotion inference engine

7.2.3.2 Image emotion inference engine

Image emotion inference engine uses the neural network trained in the trainer for image analysis. This engine analyses image data information as input and generates emotion, state vector result values.

– Input: pre-processed image data

– Output: state vector, emotion

7.3 Emotion analysis neural network – multimodal network

Multimodal network consists of a state vector integration module, a multimodal trainer module, and a multimodal emotion inference engine as presented in Figure 6.

The state vector integration module combines the state vector generated in the text analysis module, the state vector generated in the speech analysis module and the state vector generated in the image analysis module and produces an integrated state vector as the output. That is, the state vector integration module formulates the common feature representation by integrating all the high-level features extracted from the unimodal networks.

The training module performs the training process of the neural network for the multimodal emotion inference engine at the training stage.

The multimodal emotion inference engine analyses the relation among the features from the different modalities and recognizes the emotion using the neural network which consists of an input layer, at least two hidden layers and a softmax layer. The output of the softmax layer is the corresponding emotion vector. The emotion vector, for example, can be classified as "happiness."

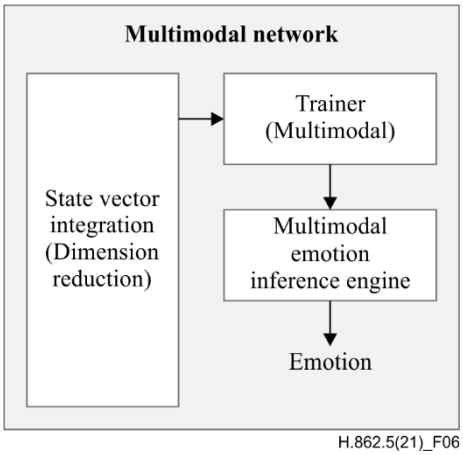

Figure 6 – Multimodal network module

7.3.1 State vector integration

The state vector integration function combines the state vectors generated from the unimodal network modules and produces the integrated state vector.

– Input: state vectors

– Output: integrated state vector

7.3.2 Trainer for multimodal network

The trainer trains the neural network with the integrated state vector.

– Input: integrated state vector

– Output: multimodal emotion inference engine

7.3.3 Multimodal emotion inference engine

Multimodal emotion inference engine analyses the integrated state vector from the input data and makes inferences on the multimodal emotion.

– Input: integrated state vector

– Output: multimodal emotion

7.4 Emotion expansion module – complex emotion generation

The emotion obtained by multimodal network expands to complex emotion through a complex emotion generation module. The complex emotion generation module, which is presented in Figure 7, decides on the final emotion based on the probabilities on the basic emotion types produced from the multimodal network.

The complex emotion generation module can decide on one specific emotion as the final emotion type in case the probability of one basic emotion is eminently higher than other emotion types (for example when the probability is higher than 80%).

When the probability of two types of basic emotion is eminently higher than other emotion types (for example when the probability of top rank basic emotion and second rank on is higher than 40%, respectively), the complex emotion generation module can decide on the complex emotion corresponding to the combination of two basic emotion based on the artificial emotion model, as the final one.

Alternatively, the emotion inference result values can be different on each unimodal case, which is when, for instance, the basic emotion type inferred from the speech is different from that of the image. In that case, the complex emotion generation module decides on the final complex emotion

corresponding to a combination of the emotion types inferred from the three different unimodal networks

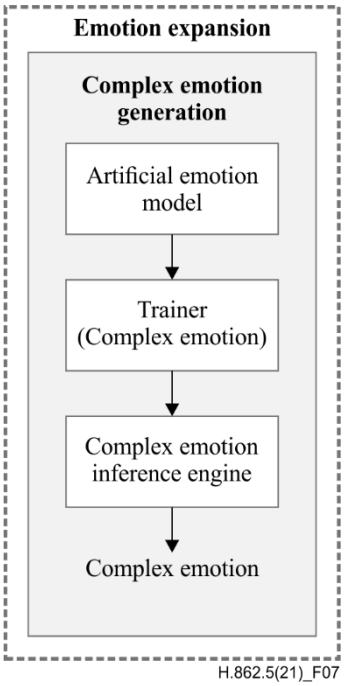

Figure 7 – Emotion expansion module

7.4.1 Artificial emotion model

The artificial emotion model generates complex emotion data by combining basic emotions based on emotion probabilities.

– Input: basic emotion data

– Output: complex emotion data

7.4.2 Trainer for complex emotion generation

The trainer trains neural networks for complex emotion generation using complex emotion data.

– Input: complex emotion data

– Output: complex emotion inference engine

7.4.3 Complex emotion inference engine

Complex emotion inference engine makes inferences on complex emotions by analysing complex emotion data and basic emotion data (multimodal emotion and unimodal emotion).

– Input: complex emotion data, multimodal emotion, unimodal emotion

– Output: complex emotion

7.5 Multimedia knowledge database

Multimedia knowledge database presented in Figure 8, generates and store datasets to train the emotion recognition model based on a large amount of multimedia data received from outside sources.

In the multimedia knowledge database, the intelligent studio tool helps to generate labelled data for given videos by annotating text-emotion data, speech-emotion data, and image-emotion data with "given emotion tags which are provided by the data manager". The labelled data can be used to train the unimodal networks and multimodal networks.

In the general training process for emotion recognition, the multimodal data are collected from the target contents which contain more than one type of media data. From the collected data, the primary feature vectors are generated for each type of media data. User data are also used for generating the primary feature vectors based on the user information and the content information of the user's preferences. These two types of primary feature vectors are integrated to generate the secondary feature vectors, which will then be used in deciding the corresponding emotion vectors based on the multiple training data. In the multimodal emotion recognition process, the relations between the resulting multiple secondary feature vectors from different media types and the emotion vectors are used as the training data set to recognize emotion.

The training processes for different media types are explained in the following clauses.

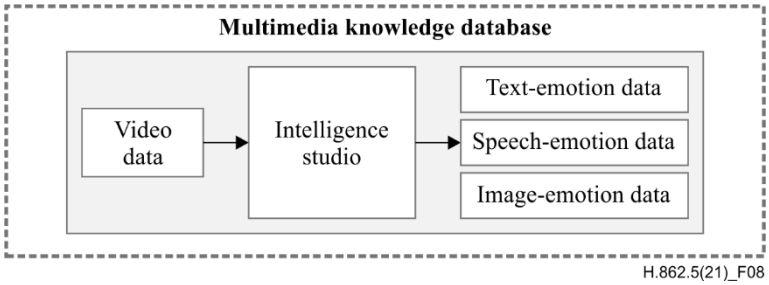

Figure 8 – Multimedia knowledge database module

8 Application to multimodal emotion analysis

When people talk, they use multiple modalities. Emotion is one of the key features to understand the meaning of the utterances made by the speaker. Therefore, a conversation system with the capability to recognize emotion can better understand the user and produce a better reply. It is a human-machine conversation system where the computer can recognize emotion in the user's speech and/or text, also using the video information of the face of the human to produce a reply.

In the conversation system with emotion capability, emotion is recognised in the following way typically and reflected in the speech production side. First, a set of emotion related cues is extracted from text, voice, and video. Then, each recognition module for text, voice, and video, recognises emotion independently. The emotion recognition module determines the final emotion based on each emotion. The emotion will be transferred to the dialogue processing module. Then the dialogue processing module produces the reply based on the final emotion and meaning from the text and video analysis. Finally, the speech synthesis module produces the speech from the reply through text. The production side can also produce the reply with an emotion.

A typical application that uses emotion in the conversation is the chatbot application. The chatbot application is widely used to provide services such as answering questions for customers in different industry domains. The multimodal capabilities enhance the customers' satisfaction with the provided services. The emotion analysis function also provides a better understanding of the customers' needs. For example, a user who uses the virtual customer service may also be able to receive a service with a smile from a chatbot. The user can interact with the chatbot via a multimodal interface such as text, voice, image, music, facial expression, video, etc. The modalities and media types used can vary depending on the application scenario.

Other example applications that use multimodal emotion analysis include a service where speech carries information not only about the lexical content, but also about a variety of other aspects such as age, gender, signature, and the emotional state of the speaker. Speech synthesis is evolving towards supporting these aspects. This is the case, for instance, of a human-machine dialogue where the message conveyed by the machine is more effective if it carries an emotion properly related to the emotion detected in the human speaker.

Another example application is in the machine translation service domains. Automatic speech translation technology denotes technology that recognizes a voice uttered in a language by a speaker, converts the recognized voice into another language through automatic translation, and outputs a converted voice as text-type subtitles or as a synthesized voice thereby preserving the speaker's features including their emotion features in the translated speech. As interest in voice synthesis increases among main technologies for automatic interpretation, the research is concentrated on personalized voice synthesis, which is a technology that outputs a target language as a synthesis voice which is similar to the tone (or an utterance style with emotion) of a speaker.

Figure 9 shows an example of a personalized automatic translation system that reflects emotional features in the translated speech. The automatic interpretation system for generating a synthetic sound having characteristics similar to those of the original speaker's voice including a speech recognition to generate text data for an original speech signal of the original speaker and extract characteristic information such as the pitch information, vocal intensity information, speech speed information, emotion features and vocal tract characteristic information of the original speech. The text data is then produced by the speech recognition module and goes through the automatic translation module to generate a synthesis-target translation and a speech synthesis module to generate a synthetic sound that resembles the original speaker using the extracted characteristic information.

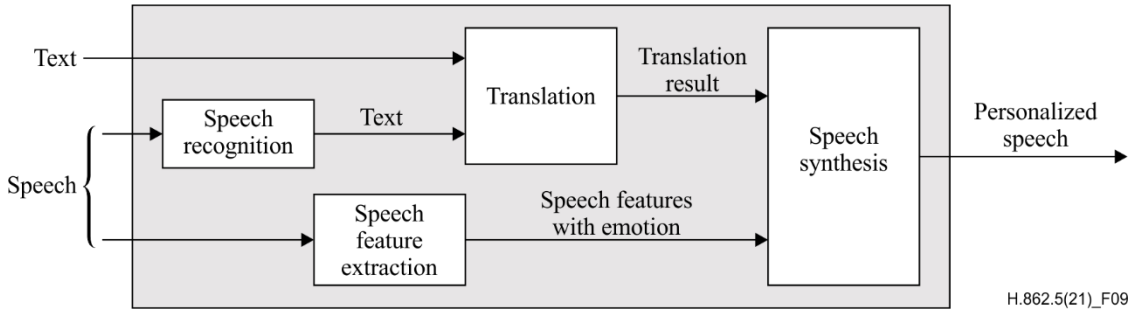

Figure 9 – Personalized automatic translation system

Appendix I

Survey on SDO's activity regarding typical emotional service

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

I.1 SDO activities on emotion

Emotion-based studies are already progressing in several SDOs and is briefly summarized in some ITU-T Recommendations that are published. However, before further adopting new current neural networks in emotional-based multimodality user interfaces (UI), it is necessary to review the status of these activities.

I.1.1 W3C

The multimodal interaction working group [b-W3C MMI] has published a working draft of EMMA, *Extensible MultiModal Annotation markup language* Version 2.0 [b-W3C EMMA]. This specification describes markup for representing interpretations of user input (speech, keystrokes, pen input, etc.) and productions of system output together with annotations for confidence scores, timestamps, medium, etc. It forms part of the proposals for the W3C multimodal interaction framework.

The emotion incubator group will discuss and propose scientifically valid representations of those aspects of emotional states that appear to be relevant for several use cases. The group will condense these considerations into a formal draft specification.

Emotion-oriented (or "affective") computing is gaining importance as interactive technological systems become more sophisticated. Representing the emotional states of a user or the emotional states to be simulated by a user interface requires a suitable representation format. Although several non-standard mark-up languages containing elements of emotion annotation have been proposed, none of these languages have undergone thorough scrutiny by emotion researchers, nor have they been designed for the generality of use in a broad range of application areas.

The deliverables are:

– Final Report of the (First) Emotion Incubator Group, 10 July 2007

– Emotion Markup Language: Requirements with Priorities, 13 May 2008

– Elements of an EmotionML 1.0, Final Report of the Emotion Markup Language Incubator Group, 20 November 2008

I.1.2 ISO/IEC JTC 1/SC 35

This group is the standardization in the field of user-system interfaces in information and communication technology (ICT) environments. It provides support for these interfaces to serve all users, including people having accessibility or other specific needs, with a priority of meeting the JTC 1 requirements for cultural and linguistic adaptability [b-ISO/IEC JTC 1/SC 35].

This includes:

– user interface accessibility (requirements, needs, methods, techniques and enablers);

– cultural and linguistic adaptability and accessibility (such as evaluation of cultural and linguistic adaptability of ICT products, harmonized human language equivalents, localization parameters, voice messaging menus);

– user interface objects, actions, and attributes;

– methods and technologies for controlling and navigating within systems, devices and applications in visual, auditory, tactile and other sensorial modalities (such as by voice, vision, movement, gestures);

– symbols, functionality and interactions of user interfaces (such as graphical, tactile and auditory icons, graphical symbols and other user interface elements);
– visual, auditory, tactile and other sensorial input and output devices and methods in ICT environments (for devices such as keyboards, displays, mice);
– user interfaces for mobile devices, handheld devices and remote interactions

Further, ISO/IEC CD 30150 is under development.

– Information technology – Affective computing user interface – Framework

I.1.3 OMA

OMA multimodal and multi-device service requirements' document describe the requirements for multimodal and multi-devices services in the scope of the OMA architecture. Such services enable access to mobile services through different modalities (e.g., keypad, graphic user interface (GUI), voice, handwriting) or devices.

– The requirements are primarily viewed from an end user point of view. Implications are derived in terms of requirements on network operators, service providers and terminal manufacturers.
– The requirements apply only to the supporting applications or services for which multimodal or multi-device interactions make sense and are desired. Applications and services may otherwise be designed without providing such a user experience.

OMA multimodal and multi-device enabler architecture documents describe the architecture needed for the OMA multimodal and multi-device enabler. Such a multimodal and multi-device enabler enables access to mobile services through different modalities (e.g., keypad, graphic user interface, voice, handwriting) or devices.

– The architecture enables applications or services supporting multimodal or multi-device user interactions with minimal changes to the programming model for non-multimodal and multi-device applications or services such as browsing.

Bibliography

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

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

[b-ITU-T F.746.7] Recommendation ITU-T F.746.7 (2018), *Metadata for an intelligent question answering service*.

[b-ISO/IEC JTC 1/SC 35] ISO/IEC JTC 1/SC 35, *User interfaces.* https://www.iso.org/committee/45382.html

[b-Ekman] Ekman, P. (1999). *Basic Emotions*. In T. Dalgleish and T. Power (Eds.) The Handbook of Cognition and Emotion Pp. 45-60. Sussex, U.K.: John Wiley & Sons, Ltd.

[b-W3C MMI] Multimodal Interaction Working Group (2015). https://www.w3.org/2002/mmi/

[b-W3C EMMA] W3C Working Group Note (2017), *Extensible MultiModal Annotation markup language Version 2.0.* https://www.w3.org/TR/emma20/

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, 2021