INTERNATIONAL TELECOMMUNICATION UNION

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

Series H

Supplement 1

(05/99)

SERIES H: AUDIOVISUAL AND MULTIMEDIA SYSTEMS

Application profile – Sign language and lip-reading real-time conversation using low bit-rate video communication

ITU-T H-series Recommendations – Supplement 1

(Previously CCITT Recommendations)

ITU-T H-SERIES RECOMMENDATIONS

AUDIOVISUAL AND MULTIMEDIA SYSTEMS

Characteristics of transmission channels used for other than telephone purposes	H.10–H.19
Use of telephone-type circuits for voice-frequency telegraphy	H.20–H.29
Telephone circuits or cables used for various types of telegraph transmission or simultaneous transmission	H.30–H.39
Telephone-type circuits used for facsimile telegraphy	H.40–H.49
Characteristics of data signals	H.50–H.99
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.399
Supplementary services for multimedia	H.450–H.499

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

SUPPLEMENT 1 TO ITU-T H-SERIES RECOMMENDATIONS

APPLICATION PROFILE – SIGN LANGUAGE AND LIP-READING REAL-TIME CONVERSATION USING LOW BIT-RATE VIDEO COMMUNICATION

Summary

Sign language and lip-reading are two important application areas of video communication. For the successful transmission of the components of visual language, certain quality requirements must be met.

This supplement is an application profile document that gives the background to the requirements and offers as well guidance on how these requirements can be met. The purpose of this supplement is not to propose new video coding schemes, but rather to indicate how current and future video coding schemes can be applied to these two areas of application, with good results.

Source

Supplement 1 to ITU-T H-series Recommendations was prepared by ITU-T Study Group 16 (1997-2000) and was approved under the WTSC Resolution No. 5 procedure on 27 May 1999.

FOREWORD

ITU (International Telecommunication Union) is the United Nations Specialized Agency in the field of telecommunications. The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of the ITU. The 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 Conference (WTSC), which meets every four years, establishes the topics for study by the ITU-T Study Groups which, in their turn, produce Recommendations on these topics.

The approval of Recommendations by the Members of the ITU-T is covered by the procedure laid down in WTSC Resolution No. 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 publication the term *recognized operating agency (ROA)* includes any individual, company, corporation or governmental organization that operates a public correspondence service. The terms *Administration, ROA* and *public correspondence* are defined in the *Constitution of the ITU (Geneva, 1992)*.

INTELLECTUAL PROPERTY RIGHTS

The ITU draws attention to the possibility that the practice or implementation of this publication may involve the use of a claimed Intellectual Property Right. The 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 publication development process.

As of the date of approval of this publication, the ITU had received notice of intellectual property, protected by patents, which may be required to implement this publication. However, implementors are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database.

CONTENTS

Included CD-ROM: "Irene" video clip

Introduction

Millions of deaf people use a sign language as their first language and are eager to use sign language for long distance conversation. The conversation speed of sign language is comparable to the speed of voice conversation.

People with varying degrees of hearing loss are considerably aided in "perceiving" speech by viewing the face of the speaker for lip-reading.

This supplement describes the importance of different factors that should be taken into account in the application of low bit-rate video coding for acceptable use in sign language and lip-reading.

The requirements set out in this supplement have been drawn from user experience but should not, however, be taken as fixed and absolute values. Different situations may call for either more stringent, or more relaxed, requirements.

Supplement 1 to H-series Recommendations

APPLICATION PROFILE – SIGN LANGUAGE AND LIP-READING REAL-TIME CONVERSATION USING LOW BIT-RATE VIDEO COMMUNICATION

(Geneva, 1999)

1 Scope

This application profile for sign language and lip-reading, lists the characteristics required of a video communication system for person-to-person conversation, using sign language and lip-reading, with or without audible speech.

It sets out performance requirements that should be met to ensure successful conversation.

It describes how sign language and lip-reading performance can be evaluated.

It suggests factors to be considered which are external to the video coding protocol regarding, for example, terminal design and also the environment in which terminals are used for sign language and lip-reading.

This supplement includes the "Irene" test sequence for evaluation of video communication for sign language.

2 Abbreviations

This supplement uses the following abbreviations:

CIF	Common Interchange Format (352 × 288 pixels)
fps	frames per second; pictures per second
QCIF	Quarter CIF (176 × 144 pixels)
SQCIF	Sub QCIF (112 × 96 pixels)

3 Definitions

This supplement defines the following term:

3.1 frame: One complete picture in video reproduction is called a "frame". In some systems, the frames are built by two half-images, each containing half of the information in the frames. These half-images are called "fields".

4 References

[1] HELLSTRÖM, DELEVERT, REVELIUS: Quality requirements on Videotelephony for Sign Language, *Swedish National Association of the Deaf*, 1997.

[2] ITU-T Recommendation G.114 (1996), *One-way transmission time*.

[3] FROWEIN: Improved speech reception through videotelephony, *IEEE journal on Selected Areas in Communication*, May 1991.

[4] ITU-T Recommendation P.931 (1998), *Multimedia communications delays, synchronization and frame rate measurement*.

5 Basic needs for reproduction of sign language and lip-reading

5.1 Basic characteristics

The components of sign language are the movements and positions of the hands, eyes, mouth, face and body.

In lip-reading, the components are the movements of the face. Often lip-reading is supported by voice. In other cases it is used together with sign language. There are also profoundly deaf people who do not sign and who rely totally on lip-reading conversations.

In video-coding terms, the scene with one signer or speaker may be regarded as containing a medium to high motion content.

5.2 Temporal resolution requirements

Both sign language and lip-reading requires good visual reproduction of movements. Assuming that a system reproduces the movements with evenly distributed pictures, the following has been observed:

- Usability for sign language and lip-reading is reported to be good at 20 frames per second (fps) [1], [3].
- With some constraints, it is possible to use a frame rate from 12 fps and higher [1].
- For lip-reading, a steep increase in usability is found at increasing frame rate up to 15 fps. After 15 fps the increase in usability continues but is less pronounced. [3]
- Very limited usability is found in frame rates between 8 and 12 fps, with severe degradation in perception or speed.
- Under 8 fps, there is no practical usefulness for either lip-reading or sign language.

5.2.1 Finger-spelling

Requirements for the temporal resolution of sign language can be found, for example, in the case of finger-spelling. Finger-spelling is a technique where each letter of the alphabet corresponds to a unique hand position. Finger-spelling positions vary from country to country. Spelling is done by showing these positions in rapid sequence to form words. The words that are spelled are usually names and other proper nouns that broader signs of sign language do not cover. Finger-spelling is very rapid and often uses up to 10 letters per second. For reliable reproduction, at least two pictures per letter should be reproduced. In other words, it can be concluded that legible reproduction of finger-spelling requires at least 20 frames per second.

5.2.2 General signing

Finger-spelling is only one part of sign language. The greater part of sign language is done with signs for complete concepts, partial sentences, grammar and ordinary nouns. There are many sign languages in the world. Even if they differ, the common concepts are close enough for the reasoning in this supplement to be valid for them all. Also, during such general signing, rapid hand movements occur and short blinks of the eyes carry grammatical information. In many cases, the temporal resolution requirements are similar to those needed for finger-spelling.

5.2.3 Lip-reading

A raw requirement figure for lip-reading can be calculated from the phoneme rate of normal speech. A normal rate is 10 phonemes per second. At least 20 pictures per second should be reproduced to enable the viewer to read the visible phonemes.

5.2.4 Adaptation

In both the case of lip-reading and sign language, the speed of language production can be slightly reduced at will. That explains why it would be possible to use 12-15 frames per second at certain times. Experienced lip-readers and sign language users also have the advantage of guessing from past experience and from redundancy. Thus it is possible for some users to have short conversations over connections lower in quality than the requirements above indicate.

5.2.5 Analysis of the frame rate requirement

An analysis of the test sequence "Irene" explains the needs further.

Finger-spelling

Table 1 shows an approximate representation of a finger-spelling sequence in the test sequence "Irene". The pictures from this sequence are shown in Figure 1.

Table 1 – Example of finger-spelling representation in frames at 25 and 12.5 frames per second

Frame No.	308		310					315					320					325					330					335	336
25 fps	**e**	**e**	**e**	**–**	**d**	**s**	**s**	**s**	**s**	**–**	**v**	**v**	**v**	**–**	**i**	**–**	**–**	**k**	**k**	**k**	**–**	**e**	**n**	**n**	**n**	**n**	**n**	**n**	**n**
12.5 fps		**e**		**–**		**s**		**s**		**–**		**v**		**–**		**–**		**k**		**k**		**e**		**n**		**n**		**n**	

The numbers in the upper row are frame numbers from the beginning of the sequence. The letters indicate when the letters are quite clearly formed by the hand. A dash indicates that no clear letter is formed in the transition between letters. The word is "Edsviken", the name of a place.

Among these eight letters, three are clearly seen only on one frame and would therefore risk being lost at 12.5 frames per second. That frame rate appears if every second frame was skipped in the coding scheme. An example of a 12.5 fps sampling is also given in the lower row of the table. It shows that only "Esvken" remains from the original word "Edsviken". This clearly demonstrates the risk of loss of language content when the frame rate is lower than 20 fps.

The distribution of letters in the 25 fps sequence is:

1 frame 3 letters;

2 frames 0 letter;

3 frames 3 letters;

4 frames 1 letter;

7 frames 1 letter (ending of phrase).

Mean length inside phrases: 2.3 frames per letter.

Conclusion

In this example, the letters within words vary between 1 and 4 frames in time, with frames representing 40 ms each. The mean length is 2.3 frames visibility per letter. The example is not long enough to make any real statistical conclusions. However, it can be seen that, with this finger-spelling speed, a frame rate of 25 fps would be sufficient while 12.5 fps would require some guesswork to perceive the finger-spelled words.

Figure 1 – The frames containing the finger-spelled word "Edsviken" recorded at 25 frames per second

General signing

Large parts of the film clip "Irene" are signed with signs without finger-spelling.

A simple analysis has been performed on one phrase. The phrase is presented here. It is transcribed sign by sign with the number of frames each sign occupies in parenthesis.

The sequence is found between frames 406 and 529 in the "Irene" sequence.

"SHE(7) TELLS(7) SELF(11) HOW(4) SHE(2) FELT(11) EXPERIENCED(13) ADOLESCENCE(16)."

None of the signs in this sequence were shorter than 2 frames and none contained more rapid motion than the finger-spelling. Some signs contain larger motions and therefore impose different requirements on video coding.

Grammatical markers

There is a grammatical blink contained in frames 394 and 395 that indicates a sentence marker.

This marker is of a length sufficient for sampling at 12.5 fps.

5.2.6 Granularity of temporal resolution

In most cases, a video camera is used for video communication that follows general video standards. This means that they deliver 25 or 30 frames per second. This fact introduces a granularity in the useful frame-rates. In using such cameras, there is not much point in talking about frame-rates between 12.5 and 25 fps or between 15 and 30 fps. Such intermediate frame-rates means that the source picture intervals will vary between 40 and 80 ms or between 33 and 66 ms respectively which introduces the risk of missing certain motion details. The conclusion is that, in order to satisfy the requirement for 20 frames per second with common cameras, the target frame-rate should be 25 or 30 frames per second.

5.3 Spatial resolution requirements

For spatial resolution it is reported that, for person-to-person sign language calls, the following is necessary: [1]

- It is possible to use QCIF resolution, but the smallest details, showing eye gaze directions, are lost. This causes stress for the recipient.
- CIF is good. The increase from QCIF to CIF gives better language perception.
- SQCIF is too coarse for reliable perception, while some signs can occasionally be perceived.
- If different resolutions are used for different parts of the picture, the hands and the face would require the highest resolution. In such schemes, care must be taken not to introduce distortion in other parts of the picture that may distract the user.

A simple theoretical verification can be done. In the head to stomach view, usually used in person-to-person sign conversation, a finger is approximately 1/50 of the picture width. In order to resolve finger(s) reliably in a picture, a finger should be represented by at least 3 pixels. That puts the minimum spatial resolution requirement to QCIF, that contains 176 pixels in width. Eye gaze direction is also important in visual language and requires higher resolution. Therefore CIF is better, and preferred.

For lip-reading, the view in person-to-person calls can be reduced to slightly more than the head. In this case, QCIF is reported to have sufficient resolution for lip-reading. [3] When using QCIF resolution, the terminal user must make sure that the display is viewed at a suitable distance so that the relatively low resolution does not cause extra disturbed perception.

5.4 Fidelity

In video communication, blur often occurs during motion.

The models for describing blur are not well developed. There are great variations in type of blur and its effect on perception. Therefore, this discussion is restricted to a brief comparison of acceptable blur on different occasions.

VHS video is reported to be sufficient for good perception of sign language and lip-reading. In video recordings, rapidly moving objects are often shown with considerable blur because shutter speeds are normally 1/50 to 1/60 of a second. This indicates that blur is acceptable on rapidly moving objects involving big movements in signing.

During large movements, some blur may be introduced occasionally. The spatial resolution during such moments should never be lower than that corresponding to SQCIF. This conclusion is drawn from the fact that SQCIF is found to be too coarse for reliable perception.

For good perception, when CIF is the base spatial resolution, the occasionally introduced blur should not go beyond what is perceived at QCIF resolution.

5.5 Delay

End-to-end video delay, from the sending camera to the receiving display, is critical in the conversation application. Values below 0.4 s are preferred, with an increase in preference down to 0.1 s. [3]

Values over 0.8 s are felt to hinder a good sign conversation. [1]

It seems that the requirements for sign language and lip-reading are similar to voice conversation requirements. The time from one utterance, until the expected reaction is seen or heard, is at least twice the delay. Even the limit of 0.4 s that is specified by Recommendation G.114 [2] seems long, when it is clear that it means delaying a response by 0.8 s.

5.6 Synchronism

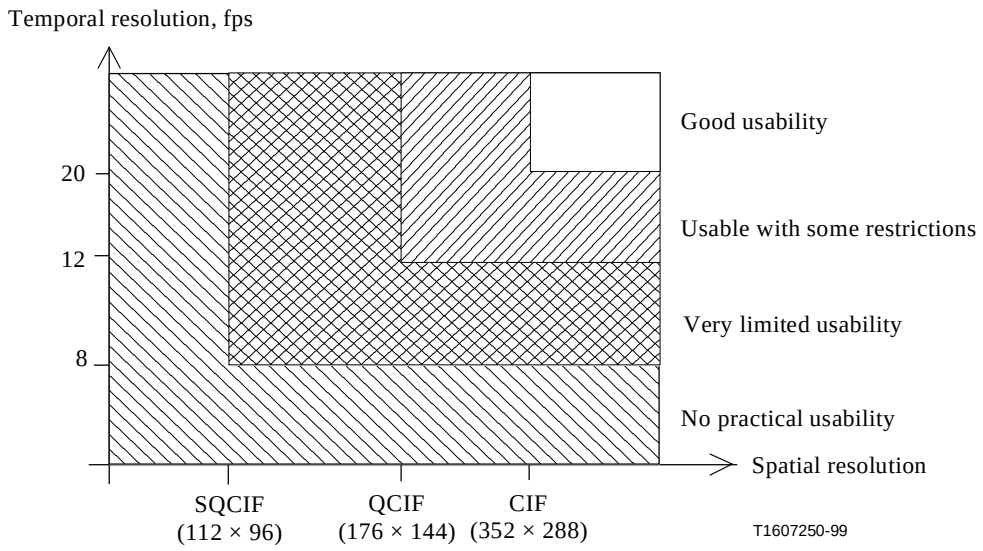
For voice-supported lip-reading, the synchronism between sound and video is essential. Time differences of up to 100 ms are reported to be acceptable. [1]

For people who can use both voice and lip-reading, the combination can be very effective for perception. [3]

5.7 Conclusion on performance requirements

For the application of sign language and lip-reading transmission in a person-to-person conversation, the following basic performance goals apply:

- Aim at 25-30 frames per second at CIF resolution and a max. 0.4-s delay, accepting occasional blur less than that corresponding to QCIF during medium motion.
- Accept, if needed in very low bit-rate environments, 12-15 fps QCIF with medium motion and occasional degradation corresponding to SQCIF during large sign language motion.
- Keep sound synchronism better than 100 ms.
- End-to-end delay should be below 0.4 s. Accept, where unavoidable, up to 0.8 s.



NOTE – The values must be observed with sign language or lip-reading movements present.

Figure 2 – Resolution requirements for sign language and lip-reading in person-to-person conversation

Table 2 – Summary of usability degradation caused by delay and blur

Usability	End-to-end delay	Occasional blur during large motions	
		For CIF resolution	For QCIF resolution
Good	<0.4 s	No	–
Usable with some restrictions	0.4 – 0.8 s	Degrade to ≅ QCIF	No
Limited usability	0.8 – 1.2 s	Degrade to ≅ SQCIF	degrade to ≅ SQCIF
No practical usability	>1.2 s	Degrade to < SQCIF	degrade to < SQCIF

6 Performance verification

6.1 Reference material

This supplement includes a CD-ROM which contains a sign language video clip that can be used for performance evaluation. The video clip "Irene" is taken from a Swedish TV programme. It contains a suitable amount of motion in the sign language. It also shows the normal rapidity of motion.

This clip is under copyright from the Swedish Educational Broadcasting Company. Appendix I reproduces the CD-ROM `Readme` file, which contains the copyright statement and the technical description of the electronic files.

6.2 Performance evaluations

A codec, or a terminal setup, is tested by transmitting the evaluation scene through a codec, or through a set of video phones, with a network connection. The result is recorded and evaluated. Recommendation P.931 [4] specifies an evaluation method.

The frame rate during signing is evaluated.

The selected static resolution is noted.

Any extra blur introduced during medium motion is measured by comparing the recorded frames with pictures from the same scene with resolution reduced to QCIF and SQCIF. Blur is only evaluated on hands and face.

The delay is measured.

The synchronism of audio (voice), and video (lip movements), is measured.

From these recordings the performance can be evaluated and compared to the goals described above.

For approximate evaluation of these values, when laboratory equipment is not available, a simple evaluation tool from the National Swedish Association of the Deaf can be used.

7 Advice to the terminal implementers

In order to satisfy user requirements, certain features should be implemented in the terminal.

- It should provide an interface to activate external alerting systems, e.g. flashing lights, pocket vibrator, watch-size vibrator or strong sound generators.
- Users may need to revert to text conversation sometimes. It is therefore advisable to implement the text conversation protocol T.140 in the terminal.
- The preference for over 20 fps and delay below 0.4 s calls for an algorithm with no frame skips to be used. A high frame rate automatically gives an opportunity to achieve a reasonable delay.
- Deviation from all quality requirements can be accepted up to 2 s after a scene shift.

8 Advice to the user

The user should arrange to use an environment with good lighting conditions and a plain background.

9 Broadening the scope

If the equipment is to be used for sign language or lip-reading application in videoconferencing, multicasting, broadcasting or information retrieval, the following facts change the requirements.

- The view is often broader, including both signing people and other objects. This indicates that usability should start at CIF spatial resolution.
- There are fewer possibilities for the user to give feedback in order to control perception by influencing the speaker or signer. Therefore, the higher frame rate from 20 fps is required.
- The delay requirements are less stringent. For broadcasting or information retrieval several seconds delay is acceptable. For conferencing, the delay requirements are similar to those for conversational use.
- The exact requirements for each application are outside the scope of this application profile.

APPENDIX I

Copyright and technical description of H-series Supplement 1 test material

This appendix reproduces the content of the `Readme.txt` *file of the CD-ROM.*

Supplement 1 to ITU-T H-series Recommendations (06/1999)
"*Irene*" video clip
Version 1.0, June 1999.

I.1 Copyright

I.2 Support

For distribution of update software, please contact:

Sales Department
ITU
Place des Nations
CH-1211 Geneve 20
SWITZERLAND
email: sales@itu.int

For reporting problems, please contact TSB helpdesk service at:

TSB Helpdesk service
ITU
Place des Nations
CH-1211 Geneve 20
SWITZERLAND
fax: +41 22 730 5853
email: tsbedh@itu.int

I.3 Details on the video sequence

This video sequence presents sign language intended to be used as test material for video coding. It contains sign language performed at natural speed.

The sequence is named "Irene" after the signing person. It is in Swedish sign language and originally produced by the Swedish Educational Broadcasting Company. It shows the same head-to-stomach view that is usually used in personal use of videophones for signing. It is recorded in PAL with 25 fps. It is provided in three formats:

1) Sign_Irene.mpeg (3261 Kbytes)

   MPEG-1 coded in CIF resolution at 25 fps;

2) Sign_Irene.cif (80 190 Kbytes)

   YCbCr 4:2:0 format in CIF at 25 fps;

3) Sign_Irene.qcif (20 048 Kbytes)

   YCbCr 4:2:0 format in QCIF at 25 fps.

Fingerspelling content

This is an approximate representation of two fingerspelling sequences in "Irene". The numbers are frame numbers from the beginning of the MPEG version. The letters indicate when the letters are quite clearly formed by the hand. A dash indicates that no clear letter is formed in the transition between letters. The first is "Pia Wickman", with the last "a" only visible on the mouth.

"Pia Wickman"

Fr.	Ltr.	Fr.	Ltr.	Fr.	Ltr.	Fr.	Ltr.	Fr.	Ltr.
29	p	39	-	49	-	59	-	69	-
30	p	40	a	50	-	60	-	70	n
31	p	41	a	51	w	61	k	71	n
32	p	42	a	52	w	62	k	72	n
33	-	43	a	53	-	63	k	73	n
34	-	44	a	54	i	64	-	74	n
35	i	45	a	55	-	65	-	75	n
36	i	46	a	56	c	66	-	76	n
37	i	47	-	57	c	67	-	77	n
38	i	48	-	58	c	68	m		

" Edsviken "

Fr.	Ltr.	Fr.	Ltr.	Fr.	Ltr.	Fr.	Ltr.	Fr.	Ltr.
308	e	315	s	322	i	329	e	336	n
309	e	316	s	323	-	330	n		
310	-	317	-	324	-	331	n		
311	-	318	v	325	k	332	n		
312	d	319	v	326	k	333	n		
313	s	320	v	327	k	334	n		
314	s	321	-	328	-	335	n		

General signing content

The last phrase in the clip is signed with signs (without finger-spelling), comparable to words.

The phrase is presented here, transcribed sign by sign, with the number of frames each sign occupies in paranthesis "SHE(7) TELLS(7) HERSELF(11) HOW(4) SHE(2) FELT(11) EXPERIENCED(13) ADOLESCENCE(16)".

The sequence is found between frames 406 and 529 in the MPEG version.

Grammatical components

The clip shows a number of eye blinks, that are typical grammatical components of sign language and which are used as sentence delimiters. They are short and, in many cases, occur only on one or two frames. There is a grammatical blink in frames 394 and 395.

ITU-T RECOMMENDATIONS SERIES

Series A	Organization of the work of the ITU-T
Series B	Means of expression: definitions, symbols, classification
Series C	General telecommunication statistics
Series D	General tariff principles
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	Transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Construction, installation and protection of cables and other elements of outside plant
Series M	TMN and network maintenance: international transmission systems, telephone circuits, telegraphy, facsimile and leased circuits
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
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 and open system communications
Series Y	Global information infrastructure and Internet protocol aspects
Series Z	Languages and general software aspects for telecommunication systems