International Telecommunication Union

ITU-R

Radiocommunication Sector of ITU

Report ITU-R BT.2207-1

(05/2011)

Accessibility to broadcasting services for persons with disabilities

BT Series

Broadcasting service (television)

Foreword

The role of the Radiocommunication Sector is to ensure the rational, equitable, efficient and economical use of the radio-frequency spectrum by all radiocommunication services, including satellite services, and carry out studies without limit of frequency range on the basis of which Recommendations are adopted.

The regulatory and policy functions of the Radiocommunication Sector are performed by World and Regional Radiocommunication Conferences and Radiocommunication Assemblies supported by Study Groups.

Policy on Intellectual Property Right (IPR)

ITU-R policy on IPR is described in the Common Patent Policy for ITU-T/ITU-R/ISO/IEC referenced in Annex 1 of Resolution ITU-R 1. Forms to be used for the submission of patent statements and licensing declarations by patent holders are available from http://www.itu.int/ITU-R/go/patents/en where the Guidelines for Implementation of the Common Patent Policy for ITU-T/ITU-R/ISO/IEC and the ITU-R patent information database can also be found.

Series of ITU-R Reports

(Also available online at http://www.itu.int/publ/R-REP/en)

Series	Title
BO	Satellite delivery
BR	Recording for production, archival and play-out; film for television
BS	Broadcasting service (sound)
BT	**Broadcasting service (television)**
F	Fixed service
M	Mobile, radiodetermination, amateur and related satellite services
P	Radiowave propagation
RA	Radio astronomy
RS	Remote sensing systems
S	Fixed-satellite service
SA	Space applications and meteorology
SF	Frequency sharing and coordination between fixed-satellite and fixed service systems
SM	Spectrum management

***Note**: This ITU-R Report was approved in English by the Study Group under the procedure detailed in Resolution ITU-R 1.*

Electronic Publication
Geneva, 2011

REPORT ITU-R BT.2207-1

Accessibility to broadcasting services for persons with disabilities[1]

(2010-2011)

TABLE OF CONTENTS

[1] This matter should be communicated to the IEC and brought to the attention of ITU-T SG 16.

Foreword

There are 650 million people with disabilities in the world today – about 10% of the world's population – and their proportion and number are growing, as humanity lives longer. A disproportionately high number of those with disabilities are in developing countries. Television, radio, and Internet are an integral part of the fabric of society, and we cannot imagine a "full life' without them. Having a disability can deny normal access to the media, and this can limit life-choices, personal independence, personal fulfilment, sense of identity, enjoyment, and social cohesion.

In considering Resolution 70 (Johannesburg, 2008) of the World Telecommunication Standardization Assembly as well as Resolution 58 (Hyderabad, 2010) of the World Telecommunication Development Conference, on access to ICT for persons with disabilities, including age-related disabilities, the ITU Plenipotentiary Conference (Guadalajara, 2010) approved Resolution 175 that instructs all three ITU sectors, inter alia, "to take account of persons with disabilities in the work of ITU, and to collaborate in adopting a comprehensive action plan in order to extend access to telecommunications/ICTs to persons with disabilities, in collaboration with external entities and bodies concerned with this subject".

The following is given in the UN Convention as an explanation of the principle of "disability". "Persons with disabilities include those who have long-term physical, mental, intellectual or sensory impairments which, in interaction with various barriers, may hinder their full and effective participation in society on an equal basis with others".

Particularly important disabilities relevant for the media include:

– **hearing** disabilities;
– **seeing** disabilities;
– **aging** disabilities;
– **cognitive** disabilities;
– lack of controllability of the man-machine interface and ease of use of the **receiver or terminal**.

However, the structure of the broadcasting system, language/writing system and culture, broadcast formats vary from one country to another and affect what kind of services may be delivered.

The Convention does not ask that infinite resources be given over to providing services for those with disabilities, but it does call for "*reasonable accommodation"* for persons with disabilities. The interpretation of this is clearly a critical issue that needs much care.

The Convention offers the following explanation of reasonable *accommodation*: "necessary and appropriate modification and adjustments not imposing a disproportionate or undue burden, where needed in a particular case, to ensure to persons with disabilities the enjoyment or exercise on an equal basis with others of all human rights and fundamental freedoms".

So, what is a proportionate burden on television, radio, and Internet to provide measures that will make it possible for those with hearing, sight, or aging disabilities to consume the same services as those without disabilities? In other words: What is "reasonable"?

Each country should establish its own accessibility programs in response to the wishes of its population with disabilities, broadcast standards, technical possibilities, resources available for investment and the management circumstances of its broadcasters.

The ITU-R may have a role to play in promoting the technical research and development that will make it possible to provide such accessible services and that will ease the burden of doing so on broadcasters, and/or in defining necessary conditions and specifications for broadcasting systems

and accessible receivers. The ITU-R also has a role to play in establishing a system for sharing worldwide the results of research and development along with information and know-how on the practical operation of accessible services.

What kind of accessible broadcast services may be introduced on what timescale depends on local conditions in each country as discussed above; the following sections are intended as examples of the kind of technology that may contribute to accessible services depending on local conditions.

1 Hearing disabilities

For television viewing, the main method of making programmes accessible is by providing optional **subtitles**.

Hearing impaired people prefer television programmes, broadcast, streamed, or downloaded which include optional subtitles in the language of the intended audience. Digital television systems have made it possible for the subtitles to be cut into the picture by a simple procedure on the remote control.

For television viewing, the secondary method of making programmes accessible is by having a **Signer "in screen"** providing a sign language version of the audio**.** This can be included permanently in the picture, or it may be possible in the future for it to be optionally cut into the picture, at the user's choice, using a broadcast multimedia system.

For radio listening, the main method of making programmes accessible is by providing data which allows display of speech on a receiver screen (**speech-to-text conversion** data).

Digital radio (audio) programmes, broadcast, streamed, or downloaded, can now include data for speech-to-text display in the receiver. A text display may also be helpful for hearing impaired people to understand the radio program.

2 Seeing disabilities

For television viewing, the main method of making programmes accessible to those with seeing disabilities is to use **"audio descriptions"**. These are audio passages which explain what is happening visually in the picture. They are provided on a second audio channel which is mixed in the receiver with the normal audio in natural pauses in dialogue. Audio descriptions are particularly effective with drama.

Audio descriptions can also be helpful to those with aging disabilities to bring to their attention things they need to notice in the picture to follow the plot fully.

3 Aging disabilities

For the elderly, it can be difficult to follow the dialogue on the radio or on television because it appears to flow too quickly. The main method of making radio programmes accessible is to adjust electronically the natural silence periods in the dialogue, and thus to make the dialogue appear to be slower.

For the aged, because human response times are slower, it can be valuable to add "audio descriptions" to television programmes which help the viewer to follow the story line (e.g. a voice says "notice the clock on the wall is at five o'clock") in the pauses in dialogue.

Radio programmes available via Internet with several speed adjustment options may help aged listeners to understand the programmes.

4 Receiver user-friendliness

Receivers should be available which have users with disabilities in mind. This can be done by the inclusion of facilities that include:

– simple and self evident controls, which operate in a similar way on all receivers;

– visual and audio guides to programme selection and choice;

– facilities for subtitle display, signer display, and audio descriptions.

It is important to note that the practicality of such features varies according to the local broadcasting system and formats, and obviously requires the cooperation of receiver manufacturers.

In Annex 1 is a report on the latest studies in Japan on technologies to improve accessibility to broadcasting services. There has been a growing interest in “universal-design products” that anyone can use with ease. And with the coming of the aging society, there will be an increasing need to develop products and services while having a good understanding of the physical characteristics of people with disabilities. The radio and television – the information devices most familiar to everyone – have become an indispensable part of daily life. A pressing issue here is how to convey broadcast information to people with disabilities. Achieving universal design in broadcasting will require a comprehensive study that examines program production techniques at the broadcasting station while also considering the ease of operating receivers, a fitting function for making viewing and listening easy for each user, etc.

In Annex 2 is a report on a study on machine translation to sign language with computer-generated (CG)-animation.

Annex 1 – Speech, captioning and multimedia browsing technologies to improve accessibility to broadcasting services

Annex 2 – Machine Translation to Sign Language with CG-animation technologies to improve accessibility to broadcasting services

Annex 1

Speech, captioning and multimedia browsing technologies to improve accessibility to broadcasting services

1 Speech rate conversion technology for elderly people

Elderly people often find that speech in contemporary broadcasts is too fast for comfortable listening. Although the use of hearing aids could be considered as one way of compensating for hearing difficulties when listening to radio or TV programs, this would not be effective for all hearing difficulties that afflict the elderly. At present, no hearing aid can effectively compensate for hearing difficulties in the face of rapid speaking. NHK considered the development of hearing assistance technology for the elderly specifically for listening to radio or television broadcasts [1].

The adaptive speech rate conversion function plays speech more slowly without overrunning the program's time slot while maintaining the quality of speech. Since a time delay would be accumulated if waveform expansion were applied evenly across speech, this technology effectively shortens non-voice intervals (that is, pauses consisting of breaths or portions with only noise). It also speeds up or slows down the rate of speech delivery to model actual utterances. Time delay is gradually eliminated while maintaining a sense of slower speech [1][4].

Slowing down the speech rate without accumulating a time delay requires an appropriate balance between contracting non-voice intervals and expanding voice intervals. Previous research investigating the relationship between a "sense of slowness" and "naturalness" reported that expanding voice intervals as much as possible was effective as long as the length of non-voice intervals was maintained at a point that minimally satisfies the need for naturalness. Such technology should also be applicable to all broadcasts, including dramas and variety shows in addition to news programs and other content that consists mostly of speech. Consideration should therefore be given to handling not just speech-based information but non-voice information as well. This can be done by first observing pitch frequency (the basic frequency of speech) and calculating its signal-to-noise (*S*/*N*) ratio with background sounds and then dynamically identifying voice and non-voice information in the context of actual program sounds.

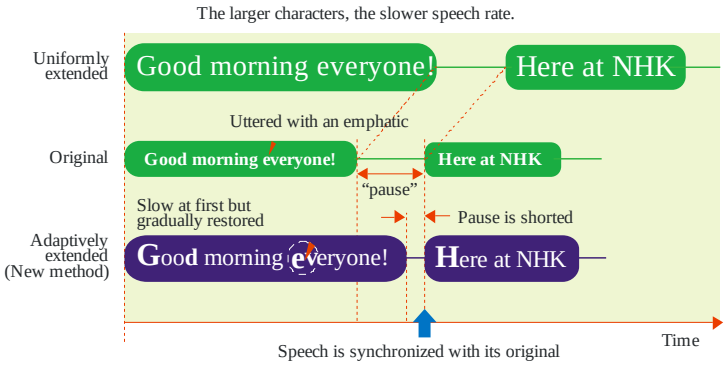
A practical speech rate conversion algorithm for incorporation in a receiver should do the following [3][4].

1. Use the *S*/*N* ratio to help identify voice intervals and non-voice intervals.
2. Allow non-voice intervals to be shortened while maintaining a time interval that does not make speech sound unnatural to the listener and allocate that deleted portion to voiced intervals.
3. Make the expansion of voice intervals variable (as opposed to uniform), placing an emphasis on expanding those portions for which an improved sense of slowness can be expected.
4. To minimize time delay accumulation, immediately suspend processing for which signal observation for longer than a certain amount of time would be required.

Based on this framework, NHK developed adaptive speech rate conversion technology as shown in Fig. 1. Here, speech that can be uttered in one breath is used as a working unit. Converted speech is then realigned with original, real-time speech after a relatively long pause (non-voice interval) corresponding to the taking of a breath. This eliminates any accumulated time delay.

FIGURE 1

Outline of adaptive speech rate conversion technology



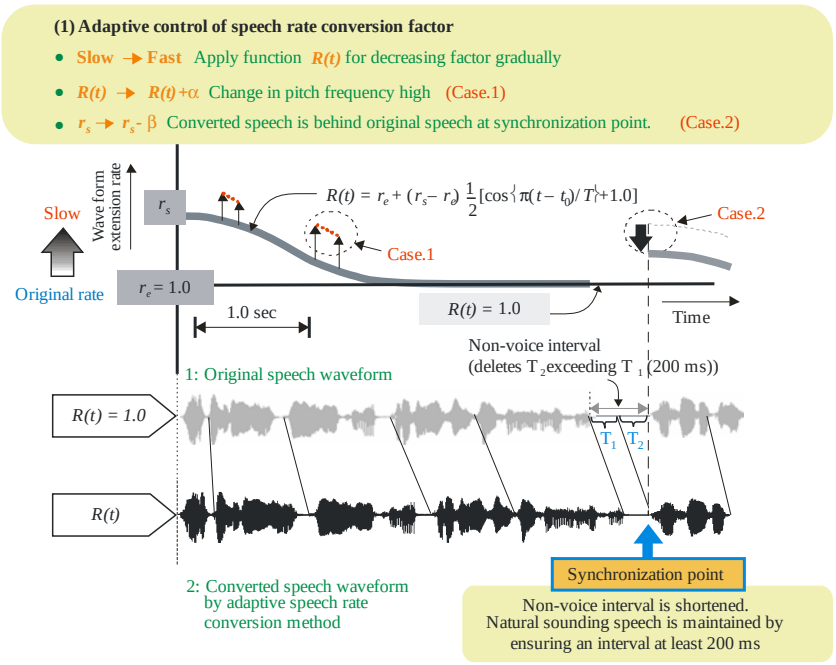
Report BT.2207-01

For radio and television announcers, it is wrong to speak on air with a uniformly slow voice; the correct way is to slow down at certain times as appropriate. In particular, a good rule of thumb is to slow down at the beginning of an utterance or during portions uttered with an emphatic, as this tends to make a good overall impression on listeners. It was therefore decided to make this rule of thumb into an engineering model. Experiments showed that the converted speech in which the initial portion was made slower was easier to listen to than the original speech of the same length [2].

The following describes the algorithm for adaptive speech rate conversion (see Fig. 2) [4].

FIGURE 2

Operation of adaptive speech rate conversion technology



Report BT.2207-02

1. In a typical intonation pattern uttered by an announcer, pitch frequency is highest in the initial portion of the utterance and falls in a nearly monotonous manner towards the end of the utterance. This gradual change is approximated by the monotonously decreasing function described below. Speech rate is changed in pace with this change in pitch frequency.

2. Figure 2 shows the method used to gradually eliminate time delay with respect to original speech. Given time period *Lp* (= 2.500 ms) as the average time taken to speak in one breath, the speech rate is gradually changed over this period (up to $T = Lp$) according to the monotonously decreasing function *R(t)*.

$$R(t) = r_e + (r_s - r_e)\ \frac{1}{2}[\cos\{\pi(t - t_0)/T\} + 1.0] \qquad (1)$$

Here, r_s and r_e are the speech rate conversion factors at the beginning and end portions, respectively, of the utterance. Their initial values are $r_s = 1.3$ and $r_e = 1.0$. When pitch frequency momentarily becomes higher, it is considered that there is some purpose behind that action and the degree of slowness at that location is temporarily increased compared to the speech rate before and after that point (Case 1).

3. If speech continues past *Lp*, speech rate conversion is generally not performed, but at $t = Lp$, r_s is reset if pitch frequency at that point in time is 70% or more of that at the beginning of the utterance.

4. If converted speech turns out to be longer than the corresponding original speech, the subsequent non-voice interval (= *T1*+*T2*) is reduced to *T1*. At about 200 ms, *T1* is the minimum time interval for which speech still sounds natural. In this case, r_s is temporarily modified downward (Case 2).

This technique prevents accumulation of time delay and enables programs to be enjoyed with slower speech even for content accompanied by video images, as in TV.

2 Real-time closed-captioning using speech recognition

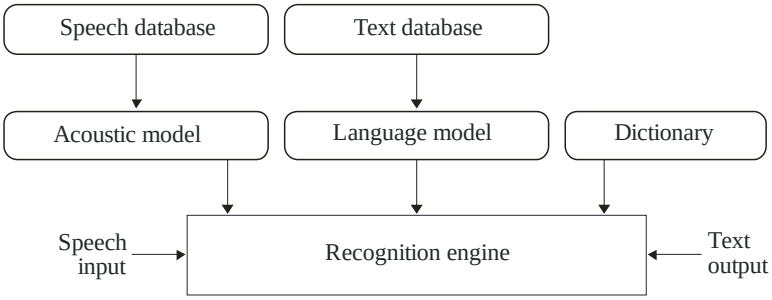
There is a great need for more TV programs to be closed-captioned to help hearing impaired and elderly people watch TV. Automatic speech recognition is expected to contribute to providing text from speech in real-time. NHK has been using speech recognition for closed-captioning of some of its news, sports, and other live TV programs [5]. In news programs, automatic speech recognition applied to anchor's speech in a studio has been used with a manual error correction system [6]. Live TV programs, such as music shows, baseball games, and soccer games, have been closed-captioned by using a re-speak method in which another speaker listens to the program and rephrases it for speech recognition [7][8].

Automatic speech recognition is a technique for obtaining text from speech using a computer. Speech recognition has greatly advanced over the last few decades along with progress in statistical methods and computers. Large-vocabulary continuous speech recognition can now be found in several applications, though it does not work as well as human perception and its target domain in each application is still limited. It has been focused on developing a better speech recognizer and applying it to closed-captioned TV programs.

A speech recognizer typically consists of an acoustic model, a language model, a dictionary and a recognition engine (Fig. 3). The acoustic model statistically represents the characteristics of human voices; i.e., the spectra and lengths of vowels and consonants. It is trained beforehand with a speech database recorded from NHK broadcasts. The language model statistically represents the frequencies of words and phrases used in the individual target domain; e.g. news, baseball or soccer. It is also trained beforehand with a text database collected from manuscripts and transcriptions of previous broadcasts. The dictionary provides phonetic pronunciation of the words in the language model. Because the recognition engine searches for the word sequence that most closely matches the input speech based on the models and the dictionary, it cannot recognize words not included in them.

FIGURE 3

Automatic speech recognizer



Report BT.2207-03

Training databases are therefore important for obtaining satisfactory speech recognizer performance. Notable features of NHK's speech recognizer are the speaker-independent acoustic model, the domain-specific language model, which is adaptable to the latest news or training texts, and the very low latency [9] from the speech input to the text output, which makes this recognizer suitable for real-time closed-captioning.

The commentaries and conversations in live TV programs such as sports are usually spontaneous and emotional, and a number of speakers sometimes speak at the same time. If such utterances are directly fed into a speech recognizer, its output will not be accurate enough for captioning because of background noise, unspecified speakers or speaking styles that do not match acoustic and language models. It is difficult to collect enough training data (audio and text) in the same domain as the target program. Therefore, the re-speak method is employed to eliminate such problems.

In the re-speak method, a speaker different from the original speakers of the target program carefully rephrases what he or she hears. This person is called the "re-speaker".

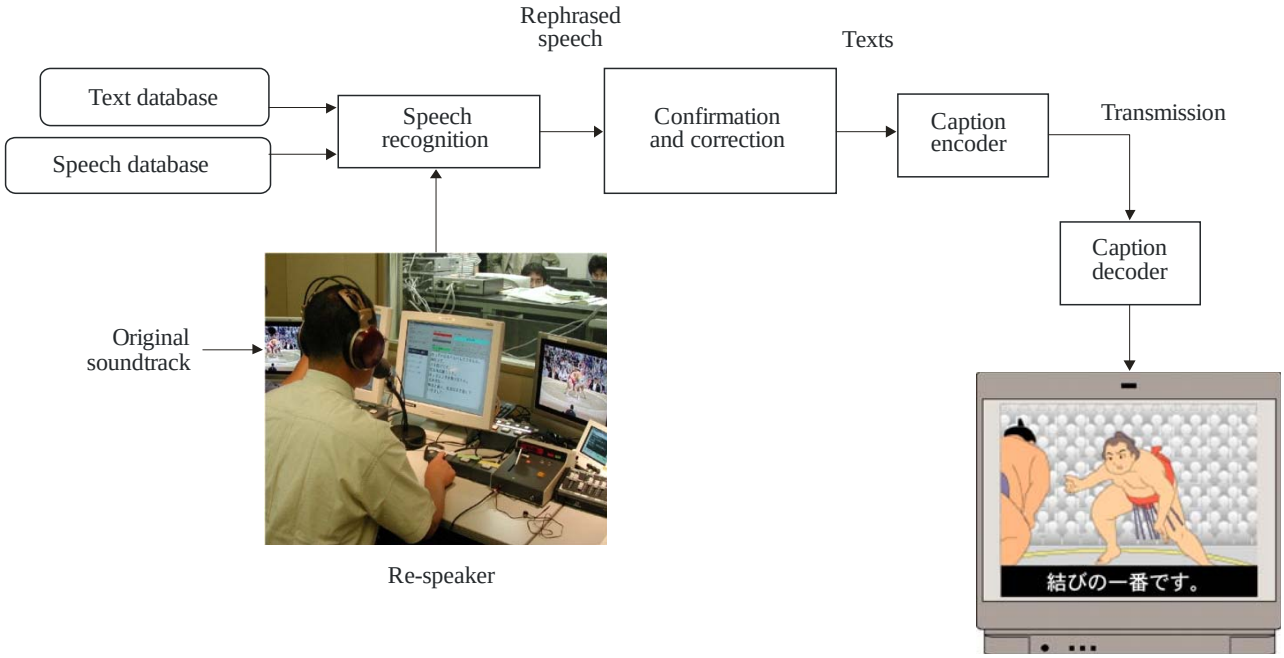
The re-speaker listens to the original soundtrack of live TV programs through headphones and repeats the contents, rephrasing if necessary, so that the meaning will be clearer or more recognizable than the original (see Fig. 4). This method provides several advantages for speech recognition.

The progress made in the speech recognition algorithms has enabled the latest speech recognizers for news programs to directly recognize not only speech read by an anchor in a studio, but also by field reporters, with sufficient word recognition accuracy of more than 95%. However, because the recognition accuracy for other parts, such as conversations and interviews, can still be insufficient, the re-speak method is still needed for those parts. Therefore, the system currently being developed is a hybrid that allows switching of the input speech for recognition between the program sound and the re-speaker's voice varying with each news item. This allows an entire news program to be covered using only the automatic speech recognizer [10].

The new speech recognizer runs on a PC. It automatically detects the gender of the speaker, which allows use of more accurate gender-dependent acoustic models [11]. As the switching of the speech input is done manually with a small delay by the re-speaker, a speech buffer of about one second is used to avoid losing any of the beginnings of utterances from the direct program sound. Moreover, the new system uses a manual correction method that requires only one or two flexible correction operators depending on the difficulty of the speech recognition. Four correction operators were needed in the previous news system (two pairs of an error pointer and an error corrector). Therefore, it is expected that the new system will help enable expansion of closed-captioned program coverage, especially for nationwide regular short news and local news programs, since their news styles are based on comparatively simple direction with only one anchor.

FIGURE 4

Closed-captioning system with a re-speak method



Report BT.2207-04

In an experiment on such simple news programs with one anchor, the new system with two correction operators achieved caption accuracy of 99.9% without any fatal errors. However, it is not yet good enough for large-scale news shows with more than one anchor and spontaneous and conversational speaking styles. Efforts are underway to improve the speech recognition accuracy for such speaking styles in the future.

3 Multimedia browsing system for visually impaired people

Integrated information barrier-free environments that will enable people with visual impairments to enjoy the wide variety of information services of digital broadcasting have been researched and developed [12][13].

Vast amounts of photographs, figures, tables and other visual content are delivered by digital broadcasting and the Internet. Moreover, the graphical user interface (GUI) enabling users to visually select news, weather, and other items of interest is by far the most pervasive way of choosing items from a menu screen. While this is convenient for people without any disabilities, it is an enormous barrier for people with visual impairments. Yet, everyone should be able to easily obtain information from the television, since virtually everyone has one at home or has access to one.

Visual impairment covers a range of disabilities from poor eyesight to total blindness, and different means of presenting information appropriate for all these degrees of disability are required. Audio presentation of information is indispensable for people who are partially sighted or blind. People with poor eyesight can comprehend information presented visually, but require an enlarged display of the items for selection. People with restricted or “tunnel” vision require the displayed items to be shown in a smaller area. In addition, people perceive brightness and colours differently, so the ability to adjust the contrast between text and background or to adjust the colours of screen items

and background is also desirable. For people who are both deaf and blind, some method of tactile input and output such as Braille or finger Braille is required.

The tactile-presentation GUI (see Fig. 5) is one method for visually disabled people to enjoy data broadcasting and to interact with the Internet [14][15]. This device forms figures and graphs on the touch display and has potential applicability as a general GUI for visually disabled people. With the touch display with optical touch panel, user can confirm details by audio or Braille, while navigating the GUI interactively and touching the visual content. The touch display has also algorithms to create real-time Braille output of menu selection buttons for data broadcasting and text data in documents.

FIGURE 5

Interactive touch display

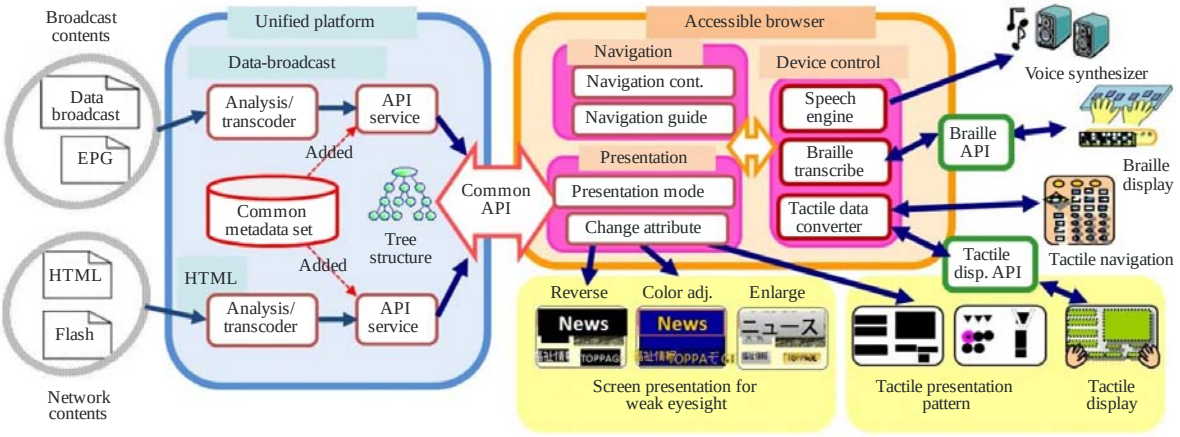
Report BT.2207-05

Figure 6 shows the configuration of the information barrier-free reception and presentation system. The browser obtains content from the integration platform via the common API. The content received is common representation data that is independent of presentation device and content format (tree structure and table structure), and semantic information is added to it in the form of metadata. The browser then converts the data to representation data for presentation by applying the various presentation methods, and it restructures the content to make it easy for the visually impaired user to understand. By modifying the attribute values of representations, the screen display can be enlarged or the colour modified to accommodate a user with poor eyesight, a list presentation can be provided for a tactile display, or other customized data can be output to devices.

In other words, the browser can present information that is tailored to the type and extent of the user's visual impairment or to the type of presentation device being used to improve access and understanding of the multimedia content [12][16].

FIGURE 6

Schematic diagram of information barrier-free reception and presentation system



Report BT.2207-06

References

[1] NAKAMURA, A., SEIYAMA, N., IMAI, A., TAKAGI, T. and MIYASAKA, E. [1996] A New Approach to Compensate Degeneration of Speech Intelligibility for Elderly Listeners. IEEE Trans. Broadcast., Vol. 42, 3.

[2] IMAI, A., SEIYAMA, N., TAKAGI, T. and MIYASAKA, E. [2001] Evaluation of Speech Rate Conversion for elderly people. Proc. of the International Workshop on Gerontechnology.

[3] IMAI, A., SEIYAMA, N., MISHIMA, T., TAKAGI, T., and MIYASAKA, E. [2001] Application of speech rate conversion technology to video editing – Allows up to 5 times normal speed playback while maintaining speech intelligibility. Proc. AES 20th International Conference, 3-5, p. 96-101.

[4] IMAI, A., TAKAGI, T., and TAKEISHI, H. [2005] Development of radio and television receiver with functions to assist hearing of elderly people. IEEE Trans. Consumer Electronics, Vol. 51, No. 1, p. 268-272.

[5] IMAI, T., HOMMA, S., KOBAYASHI, A., SATO, S., TAKAGI, T., SAITOU, K. and HARA, S. [2007] Real-Time Closed-Captioning Using Speech Recognition. ABU Technical Committee 2007 Annual Meeting, Doc. T-7/42-3.

[6] ANDO, A., IMAI, T., KOBAYASHI, A., ISONO, H., and NAKABAYASHI, K. [2000] Real-Time Transcription System for Simultaneous Subtitling of Japanese Broadcast News Programs. IEEE Transactions on Broadcasting, 46(3): 189-196.

[7] IMAI, T., MATSUI, A., HOMMA, S., KOBAYAKAWA, T., ONOE, K., SATO, S. and ANDO, A. [2002] Speech Recognition with a Re-Speak Method for Subtitling Live Broadcasts. Proc. of International Conference on Spoken Language Processing, p. 1757-1760.

[8] MARKS, M. [2003] A distributed live subtitling system. BBC R&D White Paper, WHP070.

[9] IMAI, T., KOBAYASHI, A., SATO, S., TANAKA, H. and ANDO, A. [2000] Progressive 2-Pass Decoder for Real-Time Broadcast News Captioning. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), p. 1559-1562, Vol. 3.

[10] HOMMA, S., KOBAYASHI, A., OKU, T., SATO, S., IMAI, T. and TAKAGI, T. [2008] New Real-Time Closed-Captioning System for Japanese Broadcast News Programs. 11th International Conference on Computers Helping People with Special Needs (ICCPH).

[11] IMAI, T., SATO, S., KOBAYASHI, A., ONOE, K. and HOMMA, S. [2006] Online Speech Detection and Dual-Gender Speech Recognition for Captioning Broadcast News. Proc. of the 9th International Conference on Spoken Language Processing (Interspeech 2006-ICSLP), Wed1CaP-1.

[12] SAKAI, T., HANDA, T., MATSUMURA, K., KANATSUGU, Y., HIRUMA, N. and ITO, T. [2007] Information Barrier-free Presentation System for Visually Impaired Users. CSUN Technology & Persons with Disabilities Conference 2007.

[13] HANDA, T., SAKAI, T., MATSUMURA, K., KANATSUGU, Y., HIRUMA, N. and ITO, T. [2007] Accessible EPG and Closed Caption for Visually Impaired Persons. CSUN Technology & Persons with Disabilities Conference 2007.

[14] SAKAI, T., KONDOH, S., MATSUMURA, K. and ITO, T. [2005] Improving Access to Digital Broadcasting for Visually Impaired Users", International Conference on Human-Computer Interaction 2005.

[15] HANDA, T., SAKAI, T., MATSUMURA, K., KANATSUGU, Y., HIRUMA, N. and ITO, T. [2007] An Evaluation of Accessibility of Hierarchical Data Structures in Data Broadcasting Using Tactile Interface for Visually-Impaired People. 12th International Conference on Human-Computer Interaction(HCI2007), Universal Access in HCI, Part III, HCI2007, LNCS4556, p. 45-54.

[16] MATSUMURA, K., SAKAI, T. and HANDA, T. [2007] Restoring Semantics to BML Content for Data Broadcasting Accessibility," 12th International Conference on Human-Computer Interaction (HCI2007), Universal Access in HCI, Part III, HCI2007, LNCS4556, p. 88-97.

Annex 2

Machine translation to sign language with CG-animation technologies to improve accessibility to broadcasting services

1 Machine translation to sign language with CG-animation

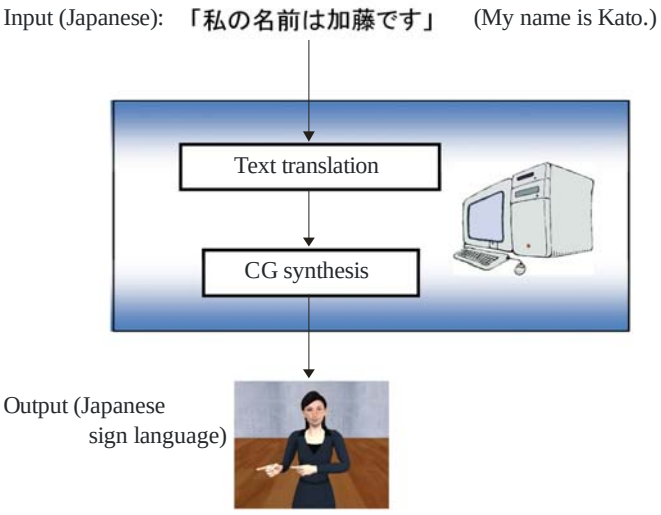
In Japan, deaf people, especially those born deaf or who lost hearing in early childhood, use Japanese Sign Language (JSL) to communicate with each other. JSL is a visual language in which words and phrases are created using not only manual signals with hand and finger gestures, but also non-manual ones with facial expressions, head movements and eye direction. These three-dimensional motions make JSL grammar different from that in spoken Japanese, which has one dimension: sound. Due to the different grammars, native signers understand JSL representations easier than spoken Japanese ones.

In broadcasting, more TV services for deaf people are needed. Closed caption service using transcription is one of such services and is widely used, partially helped with speech recognition technology. Transcription is helpful for those who became hearing impaired later in life, while it is difficult to understand for native signers, because it is based on spoken Japanese. Native signers truly need more broadcasting services with JSL, which is their mother tongue. The simplest method to increase JSL services is to increase the number of JSL translators engaging in translating TV programs from Japanese into JSL. However this is difficult to do, because Japan has too few JSL translators. Furthermore, JSL translators have to be taught to be able to translate TV programs that include a lot of jargon, and are difficult to find in the middle of the night to translate a breaking news report about an earthquake, typhoon, or so on.

To overcome these problems, NHK has been studying machine translation (MT) from Japanese to JSL with CG-animation. The MT system translates texts in Japanese into CG-animations in JSL. Figure 7 shows an overview of our goal system for MT with CG-animation.

FIGURE 7

An overview of our goal system for MT with CG-animation



Report BT.2207-07

The MT system consists of two major processes: text translation and CG synthesis. Text translation transfers words and phrases in Japanese into sequences of symbols that represent actions in JSL by using a Japanese-to-JSL dictionary, and puts these sequences into a sentence by using a set of transfer rules. CG synthesis generates seamless motion transitions between each symbol by using a motion interpolation technique and adds non-manual signals to the animation.

As the first step to realize the MT system, a Japanese-to-JSL dictionary was recently developed. Figure 8 shows an example of the online version of the bilingual dictionary.

FIGURE 8

A Japanese-to-JSL dictionary

Report BT.2207-08

The dictionary has 86.600 Japanese entries and 4.900 JSL entries with CG-animation. In the dictionary, the number of Japanese entries automatically expands to 86.600 Japanese words from 4.900 Japanese basis words corresponding to the JSL entries due to our Natural Language Processing (NLP) method, which exploits some synonyms in several lexicons to find the nearest ones in the meaning and ranks the accuracy of the synonyms for a word by using a confidence measure defined from their surface similarity and the number of the synonym lexicons in which they are registered [1]. Meanwhile, the CG-animation defines a high-quality 3D human model of hands and fingers, and controls the model using motion-capture data. The model has about 60 joints with three rotation angles and can express most of manual signals in JSL [2]. CG-animation is rendered by scripts in TVML (TV program Making Language), which is a scripting language developed by NHK to describe full TV programs [3].

A JSL corpus has also been constructed on daily NHK JSL News programs [4]. The corpus is utilized for analyzing JSL grammar and translation rules, and comparing CG-animated JSL gestures with human ones. The corpus consists of Japanese sentences, their JSL translations and their JSL videos. Figure 9 shows a browsing system for the corpus.

The Japanese sentences are transcribed by revising the speech recognition results of the news programs and their JSL translations are done by transferring the sign gestures of the newscasters to JSL letters. The JSL videos are extracted along the time intervals of the transcribed JSL translations by hand. The corpus is currently composed of about 10.000 sentences with these annotations.

A prototype system for MT from Japanese to JSL with CG-animation is under development, integrating these basic technologies and improving each module of text translation and CG synthesis, and will help deaf people to fully appreciate TV programs.

FIGURE 9

A browsing system for the JSL corpus

Report BT.2207-09

References

[1] KATO, N., KANEKO, H., INOUE, S., SHIMIZU, T. and NAGASHIMA, Y. [2009] Construction of Japanese sign language lexicon – Automatic Expansion of Japanese vocabulary. Proc. of IEICE HCG Symposium 2009, I-3 (In Japanese).

[2] KANEKO, H., HAMAGUCHI, N., DOKE, M., INOUE, S. and SHIMIZU, T. [2009] A Study of Sign Language Animation using TVML. Proc. of IEICE WIT2008-82, p. 79-83 (In Japanese).

[3] HAYASHI, M. [1998] TVML (TV program Making Language) – Automatic TV Program Generation from Text-based Script. Proc. of Siggraph 98.

[4] KATO, N. [2010] Construction of JSL News corpus. Proc. of 16th Annual Meeting of The Association for Natural Language Processing, p. 494-497 (In Japanese).

Annex 3

Audio processing technologies to improve accessibility to broadcasting services

1 Device for evaluating broadcast background sound balance for elderly listeners

Broadcast audio is produced with the aim of serving the general public, from infants to the elderly. However, it is known that the minimum audible field (MAF) generally increases with age. It is also said that the elderly perceive background sound as louder and have difficulty in understanding spoken lines and narrations. When producing broadcast programs, the background sound balance is subjectively determined by individual program production mixers who have never experienced how the elderly perceive sound. To improve this situation, NHK has developed a device that helps a mixer adjust the loudness of background sounds to an appropriate level. This research focused on degradation of hearing acuity due to aging by taking the frequency bandwidth of broadcast sound and composition of the sound source into account [1], [2].

Two factors were considered as the cause of the elderly perceiving background sound as loud or noisy. One is the fact that it is difficult for them to separate and understand the narration from background sound. This may be caused by the degradation of inner ear function as well as the deterioration of processing ability in the auditory centre. The other factor arises from a production technique, often used in broadcast programs, to enhance the mood by making background music and sound effects louder when the sequence has no narration.

Elderly listeners may overreact to such sounds. This may be caused by the recruitment phenomenon due to aging, making them sensitive to sound level changes. Perception of program sound level by elderly listeners was evaluated using loudness [3] as a parameter. In this experiment, elderly listeners' ages ranged from 60 to 72. It was found that elderly listeners become annoyed if the loudness of the background sound is more than 2.5 phon louder than the narration loudness level. It was also found that elderly listeners perceive background sounds as louder when the difference in sound levels between the background sound and narration is less than 6 phon compared with when the difference is greater than 6 phon.

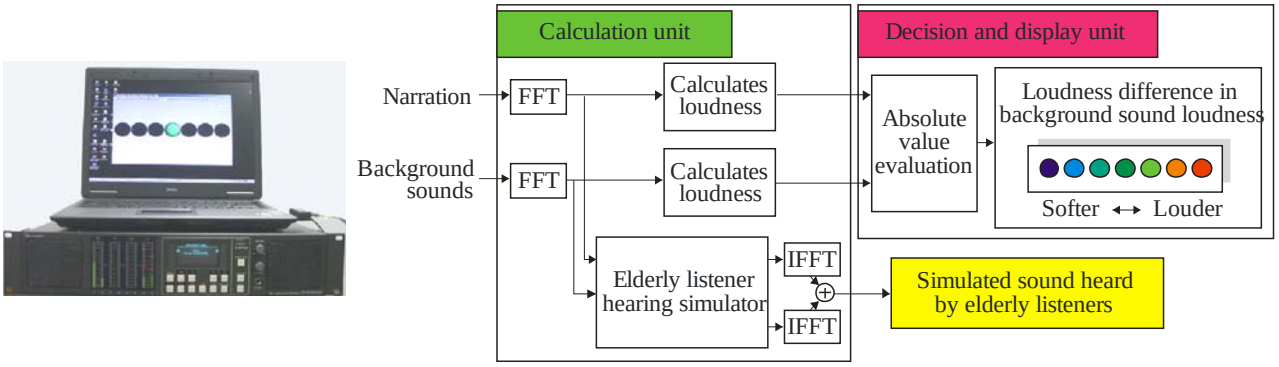
On the basis of these findings, a prototype system was developed for objectively evaluating the loudness of broadcast TV programs optimized for elderly listeners [1], [4]. Figure 10 show a

photograph and a block diagram of the system. The loudness levels of the narration and background sounds are calculated from the audio signals. The background sound balance is represented as a series of thresholds based on the loudness level, the difference in loudness level between the narration and background sound, and the listening level, and it is displayed as a seven color-coded scale (blue, aqua, blue-green, green, yellow-green, orange, red), with blue on the left signifying too soft, red on the right signifying too loud, and green in the middle signifying the optimum sound level. Figure 11 shows the threshold values for a listening loudness level of 70 phon, which is equivalent to a listening level of 75 dBA [4]. The figure compares background sound balance thresholds for people with normal hearing as determined by evaluating the program production mixers [5] and thresholds for elderly listeners. By the prototype system to enable program production staff experience the hearing difficulty of elderly people, a function to simulate deterioration of sound separation ability as well as the recruitment phenomenon is also introduced.

The prototype system was tested at a broadcasting station and found to be an extremely useful tool for aiding in the production of TV programs with the optimal sound volume balance for elderly listeners. Refinements and improvements are currently being incorporated into the prototype system.

FIGURE 10

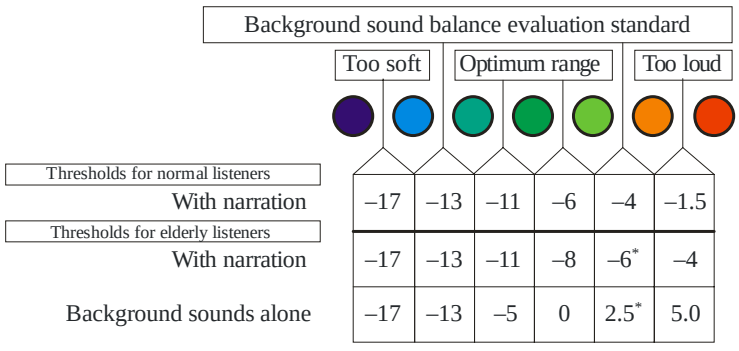
Externals and block diagram of prototype evaluation system



Report BT.2207-10

FIGURE 11

Loudness level thresholds when listening loudness level is 70 phon



Thresholds for normal listeners						
With narration	–17	–13	–11	–6	–4	–1.5
Thresholds for elderly listeners						
With narration	–17	–13	–11	–8	–6*	–4
Background sounds alone	–17	–13	–5	0	2.5*	5.0

*These red threshold values are described in this Report.

Report BT.2207-11

References

[1] KOMORI, T., TAKAGI, T., KUROZUMI, K. and MURAKAWA, K. [2008] An Investigation of Audio Balance for Elderly Listeners using Loudness as the Main Parameter, AES 125th Convention Paper 7629.

[2] KOMORI, T. and TAKAGI, T. [2009] A study of elderly people's hearing loss and the subjective evaluation result from varying background sound level of TV program, ITE Winter Annual Convention, 4-9, (in Japanese).

[3] ZWICKER, E., FASTL, H., WIDMANN, U., KURAKATA, K., KUWANO, S. and NAMBA, S. [1991] Program for calculating loudness according to DIN 45631 (ISO 532B), J. Acoust. Soc. Ja. (E)., Vol. 12, pp. 39-42.

[4] KOMORI T., TAKAGI, T., KUROZUMI, K., SHODA, K. and MURAKAWA, K. [2010] A Device to Evaluate Broadcast Background Sound Balance Using Loudness for Elderly Listeners, ICCHP 2010, Part II, LNCS6180, pp. 560-567.

[5] KOMORI, T., KOMIYAMA, S., DAN, H., TAKAGI, T., SHODA, K., KUROZUMI, K., HOSHI, H. and MURAKAWA. K. [2009] A Investigation of the Audio Balance Control based on the Loudness Level, IEICE Transactions, Vol. J92-A, No. 5, pp. 344-352, (in Japanese).