International Telecommunication Union

ITU-R

Radiocommunication Sector of ITU

Report ITU-R BT.2207-3

(10/2017)

Accessibility to broadcasting services for persons with disabilities

BT Series

Broadcasting service (television)

International Telecommunication Union

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, 2017

© ITU 2017

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without written permission of ITU.

REPORT ITU-R BT.2207-3

Accessibility to broadcasting services for persons with disabilities

(2010-2011-2012-2017)

TABLE OF CONTENTS

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 programmes 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 in the future be possible 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 that 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 programme.

2 Sight impairment

For television viewing, the main method of making programmes accessible to those with sight impairment is to use **"audio descriptions"**. These are audio passages that 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 audience

Aging audience can experience difficulties when trying 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.

It is known that through the aging process, response time tends to slow down. It can be valuable to add "audio descriptions" to television programmes in the pauses in dialogue, 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").

Radio programmes available via Internet with several speed adjustment options may help a wider age range of 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.

The Annex 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. Moreover, 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 programme 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.

Annex

Technologies to improve accessibility to broadcasting services

1 Speech rate conversion technology

As the average age of the television audience increases, broadcaster often receive comments 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 programmes, this would not be effective for all hearing difficulties especially age related hearing loss. At present, no hearing aid can effectively compensate rapid speech. The development of technologies that can assist a wide age range audience to listening to radio or television broadcasts is needed [1].

The adaptive speech rate conversion function plays speech more slowly without overrunning the programme'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 programmes 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 programme sounds.

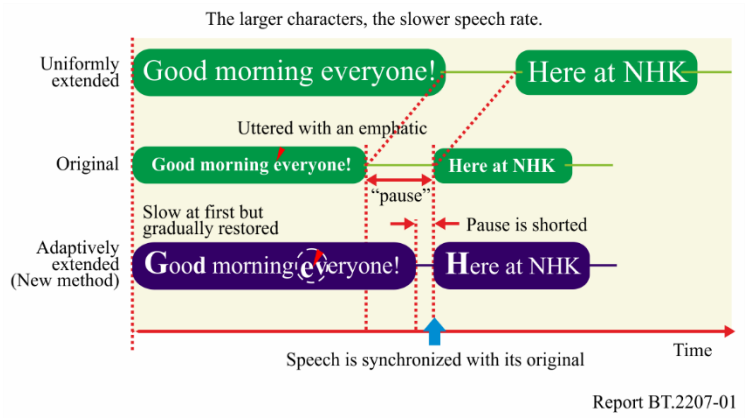
A practical speech rate conversion algorithm for incorporation in a receiver should do the following [3][4].

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

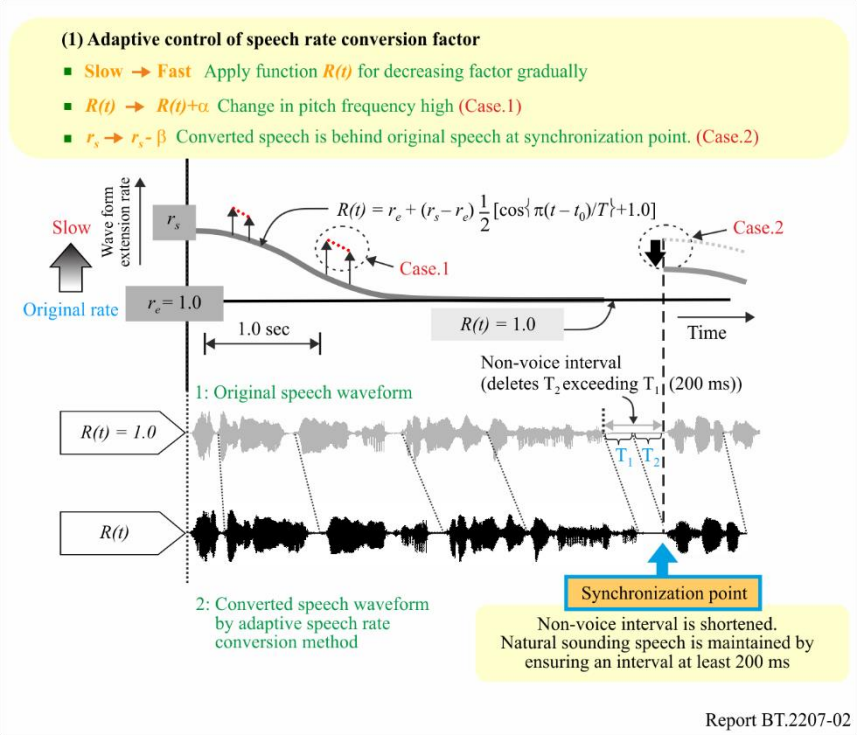


For radio and television announcers, it is wrong to speak on air with a uniformly slow voice; the correct way is to slow down at certain times as appropriate. In particular, a good rule of thumb is to slow down at the beginning of an utterance or during portions uttered with an emphatic, as this tends 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



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)*.

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

   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 programmes to be enjoyed with slower speech even for content accompanied by video images, as in TV.

This speech rate conversion technology was also implemented in the radio-on-demand services on the website. Listeners can choose from three different speech rates: 1. normal, 2. slower, and 3. faster. Normal is the original speed. Slower is 0.83 times the original speed. Faster is 1.7 times the original speed by our proposed adaptive conversion method. Figure 3 shows a page from the NHK radio news site, which is a service on the website.

FIGURE 3

Website service of radio news with speech rate conversion technology

2 Real-time closed-captioning using speech recognition

There is a great need for more TV programmes to be closed-captioned to help hearing impaired and those with age-related hearing loss to enjoy TV programmes. Automatic speech recognition is useful for producing 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 programmes [5]. In news programmes, automatic speech recognition applied to anchor's speech in a studio has been used with a manual error correction system [6]. Live TV programmes, 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 programme and rephrases it for speech recognition [7], [8].

Automatic speech recognition is a computer-based technique for creating text from speech. Speech recognition has advanced a great deal thanks to statistical analysis and increased computing power. Large-vocabulary continuous speech recognition can now be found in several applications, though it does not yet work as well as human perception and its target domain in an application is still limited. Researchers have therefore focused on developing better speech recognizers and applying them to closed-captions for TV programmes.

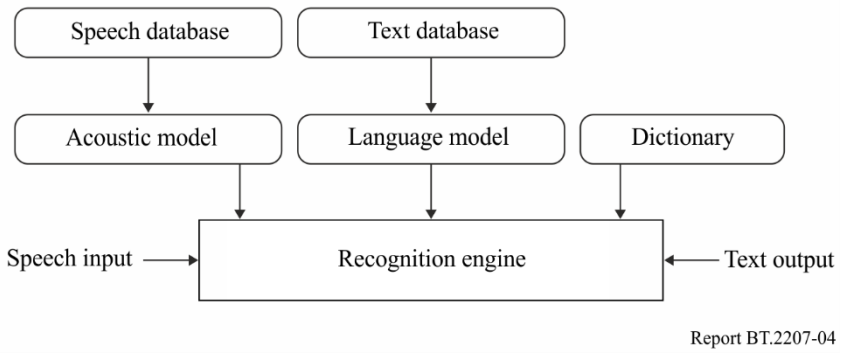
A speech recognizer typically consists of an acoustic model, a language model, a dictionary and a recognition engine (Fig. 4). The acoustic model statistically represents the characteristics of human voices; i.e. the spectra and lengths of vowels and consonants. It is trained beforehand with a database of speech 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.

Training databases are therefore important for obtaining satisfactory speech recognizer performance. The notable features of NHK's speech recognizer which make it suitable for real-time closed-captioning 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.

FIGURE 4

Automatic speech recognizer



The commentaries and conversations in live TV programmes 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 the acoustic and language models. It is difficult to collect enough training data (audio and text) in the same domain as the target programme. Therefore, the re-speak mode is employed to eliminate such problems. In the re-speak mode, a speaker different from the original speakers of the target programme 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 programmes through headphones and repeats the content, rephrasing if necessary, so that the meaning will be clearer or more recognizable than the original (see Fig. 5). This mode provides several advantages for speech recognition.

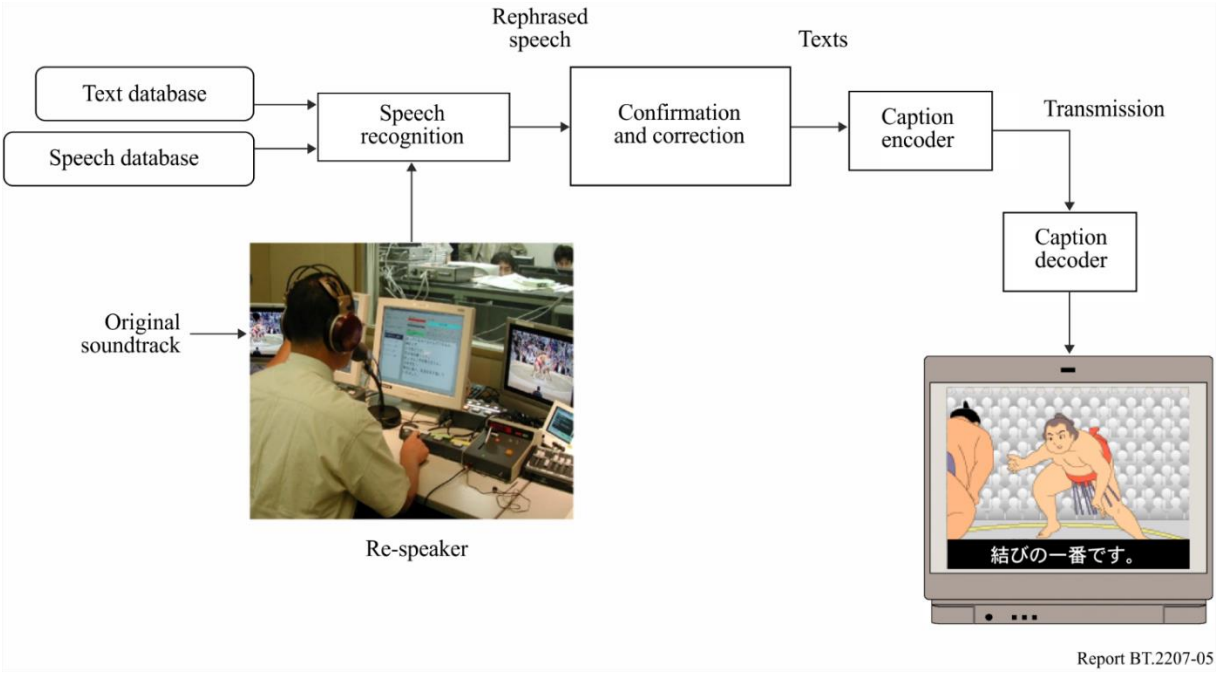
The progress made in the speech recognition algorithms has enabled the latest speech recognizers for news programmes to directly recognize not only speech read by an anchor in a studio, but also by field reporters, with sufficient word recognition accuracy (more than 95%). However, because the recognition accuracy for other parts, such as conversations and interviews, may 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 programme sound and the re-speaker's voice varying with each news item. This allows an entire news programme 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 the 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 the beginnings of utterances from the direct programme 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. Since four correction operators (two pairs of an error pointer and an error corrector) were needed in the previous news

system, the new system is also superior in terms of operating costs. In an experiment on simple news programmes with one anchor, the new system with two-correction operators achieved a caption accuracy of 99.9% without any fatal errors. The new system was initially used to caption part of news reporting on the disaster that occurred on 11 March 2011 and it has been employed to caption nationwide regular short news since March 2012. The new system will help us to expand our closed-captioned programme coverage, especially for such short news and local news programmes, since their news styles are based on comparatively simple direction with only one anchor.

FIGURE 5

Closed-captioning system with a re-speak method



However, the system still is not good enough for large-scale news shows with more than one anchor and spontaneous and conversational speaking styles. As a result, we are making efforts to improve the speech recognition accuracy for such speaking styles.

Automatic closed captioning technology based on productions that have manuscripts available in advance of transmission has been developed in the process of programme production for regional broadcasting stations. This technology compares programme manuscripts with recognized speech and instantly identifies the lines in the manuscripts corresponding to the speech. This makes it possible to eliminate manual correction when producing closed captions for broadcasting.

A news programme may be prepared using multiple manuscripts and the presentation order is not specified in advance. Often parts of the manuscripts may be skipped or rephrased. This makes it more difficult to determine the target document based on the recognition results. The identification task may also be complicated by speech recognition errors. To overcome these issues, a mechanism is incorporated a mechanism called a weighted finite-state transducer (WFST)[1], which can accurately estimate the manuscript segment that was read based on the recognized word sequence.

The system is also designed to prevent incorrect closed captions from being displayed by withholding the closed-caption display on segments that do not have supporting manuscripts, such as interviews.

[1] WFST: technology for estimating words with the highest similarity. The technology is commonly used for language processing and speech recognition.

References

[1] KOBAYASHI, A., ICHIKI, M., FUJITA, Y., OKU, T., and SATO, S.: "Discriminative Language Modeling Based on Recurrent Neural Networks," Autumn Meeting of the Acoustical Society of Japan, 2-Q-16 (2014) (in Japanese)

[2] ONOE, K., ICHIKI, M., OKU, T., KOBAYASHI, A. and SATO, S.: "Parameters of DNN for Recognition of Broadcast Speech," Spring Meeting of the Acoustical Society of Japan, 1-P-24 (2015) (in Japanese)

[3] OKU, T., ICHIKI, M., ONOE, K., KOBAYASHI, A. and SATO, S.: "Development of Speech and Language Corpora by Using Broadcast Speech and Closed Caption," The Special Interest Group Technical Reports of IPSJ, Vol. SLP-103, No. 2 (2014) (in Japanese)

[4] ICHIKI, M., ONOE, K., OKU, T., KOBAYASHI, A. and SATO, S.: "Expansion Method of Pronunciation Lexicon by Using Phoneme Translation Model Trained by Large Speech Corpus," Spring Meeting of the Acoustical Society of Japan, 1-1-9 (2015) (in Japanese)

[5] SATO, S., ONOE, K., KOBAYASHI, A., OKU, T., ICHIKI, M. and ARAI, T.: "Recent Developments in Automatic Captioning System for Regional Broadcasting Station," The Special Interest Group Technical Reports of IPSJ, Vol. SLP-103, No. 1 (2014) (in Japanese)

3 Multimedia browsing system for the visually impaired

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 that 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 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.

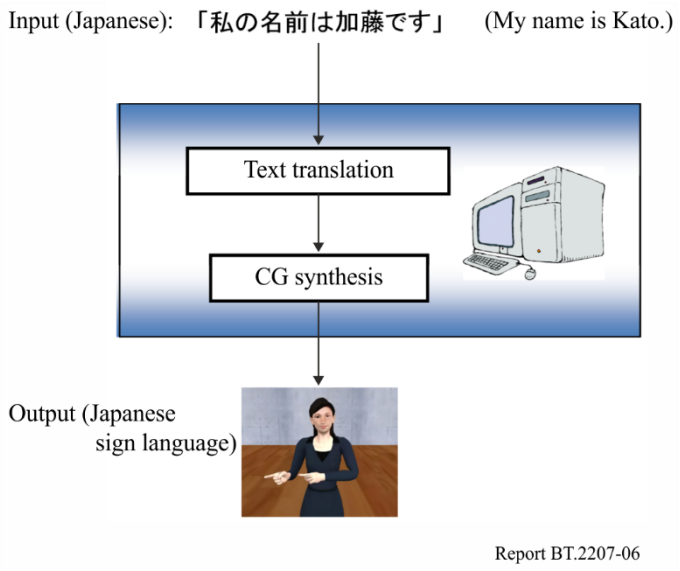
In broadcasting, more TV services for deaf people are needed. Closed caption service using transcription is one of such services and is widely used, partially helped with speech recognition technology. Transcription is helpful for those who became hearing impaired later in life, while it is difficult 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 programmes 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 programmes 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 6 shows an overview of our goal system for MT with CG-animation.

FIGURE 6

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

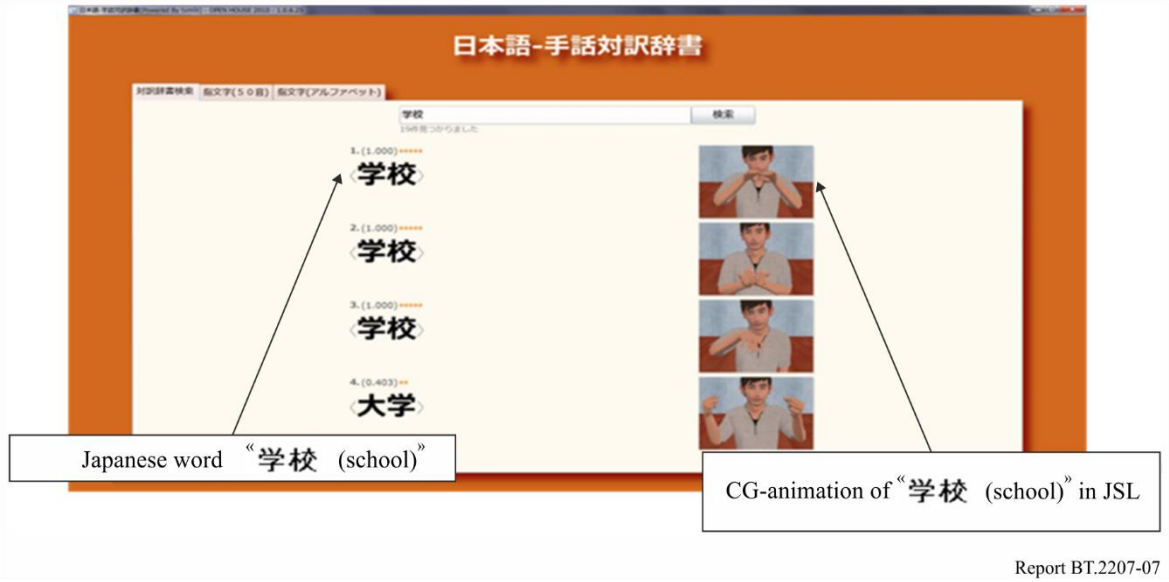


The MT system consists of two major processes: text translation and CG synthesis. Text translation transfers words and phrases in Japanese into sequences of symbols that represent actions in JSL by using a Japanese-to-JSL dictionary, and puts these sequences into a sentence by using a set of transfer 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 7 shows an example of the online version of the bilingual dictionary.

FIGURE 7

A Japanese-to-JSL dictionary



The dictionary has 100 000 Japanese entries and 7 000 JSL entries with CG-animation. In the dictionary, the number of Japanese entries automatically expands to 100 000 Japanese words from 7 000 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 programmes [3].

A JSL corpus has also been constructed on daily NHK JSL News programs [4]. The corpus is utilized for analysing 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 8 shows a browsing system for the corpus.

The Japanese sentences are transcribed by revising the speech recognition results of the news programmes 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 130.000 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 programmes.

An automatic CG sign language generation system has been developed, which converts weather forecast data provided by the Japan Meteorological Agency in XML format into CG sign language animations. The system can read coded weather forecasts and convey them through animated characters using sign language. With the help of hearing impaired volunteers and sign language interpreters, signing phrase templates and a signing motion database was created that correspond to the weather forecast codes. The system receives and analyses the XML code, then inserts the relevant information into the signing phrase templates, and then automatically generates the appropriate sign language animations (see Fig. 9).

The processing speed and stability of the system was improved by integrating processes from weather data reception and CG sign language animation generation via data upload to a website. A website for evaluating the CG sign language for weather report automatically generated by this system (see Fig. 10) was released through NHK Online in February 2017. (See https://www.nhk.or.jp/strl/sl-weather/)

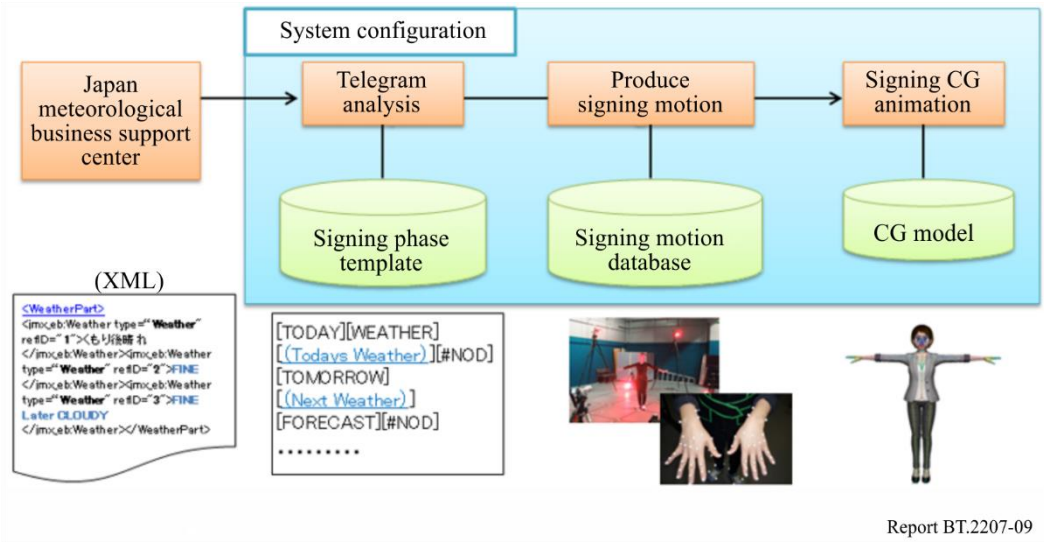
FIGURE 8

A browsing system for the JSL corpus

Report BT.2207-08

FIGURE 9

Configuration of automatic CG sign language generation system for weather forecasts



Report BT.2207-09

FIGURE 10

Website for evaluating weather forecast sign language CG (on NHK online)

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).

[5] HIRUMA, N., AZUMA, M., UCHIDA, T., UMEDA, S., MIYAZAKI, T., KATO, N., INOUE, S. [2015] Automatic Generation System of Japanese Sign Language (JSL) with CG Animation of Fixed Pattern Weather Information. ABU Technical Review, no.264, 2015, p. 2-5.

[6] AZUMA, M., HIRUMA, N., UCHIDA, T., MIYAZAKI, T., INOUE, S., UMEDA, S., KATO, N., [2016] Development of Automatic Sign Language Animation System to Express Weather Warnings. Proc of IEICE Technical Report, Vol.116, No.248, WIT2016-35, pp.11-15 (2016) (in Japanese).

5 Device for evaluating broadcast background sound balance

Broadcast audio is produced with the aim of serving the public of all ages. However, it is known that the minimum audible field (MAF) generally increases with age. It is also said that background sound seems to become louder with age, causing difficulties in understanding spoken lines and narrations. When producing broadcast programmes, the background sound balance is subjectively determined by individual programme production mixers who have excellent hearing and do not adequately appreciate the impact of age related hearing issues. 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 perceiving background sound as loud or noisy. One is the fact that with age it becomes more difficult 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 programmes, to enhance the mood by making background music and sound effects louder when the sequence has no narration.

The recruitment phenomenon due to aging may cause over reaction to background sound and increase the sensitivity to sound level changes. Perception of programme sound level by listeners in the age range 60 to 72 was evaluated using loudness [3] as a parameter. In this experiment, it was found that the 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 the 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.

Based on these findings, a prototype system was developed for objectively evaluating the loudness of broadcast TV programmes optimized for listeners in this age range [1], [4]. Figure 11 shows 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 colour-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 12 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 programme production mixers [5] and thresholds for an older age range. By the prototype system to enable programme production staff experience the impact of age related hearing difficulty, 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 programmes with the optimal sound volume balance for a wider age range of listeners. Refinements and improvements are currently being incorporated into the prototype system.

FIGURE 11

Externals and block diagram of prototype evaluation system

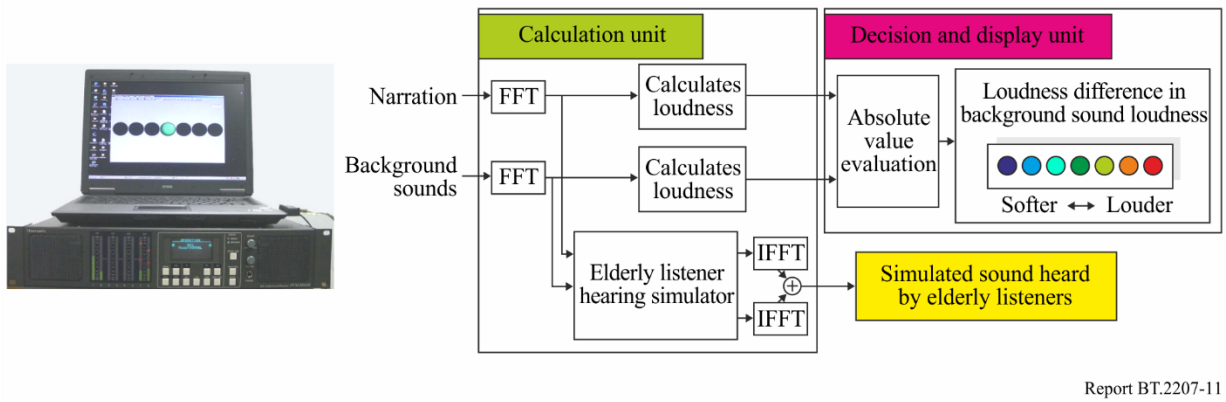
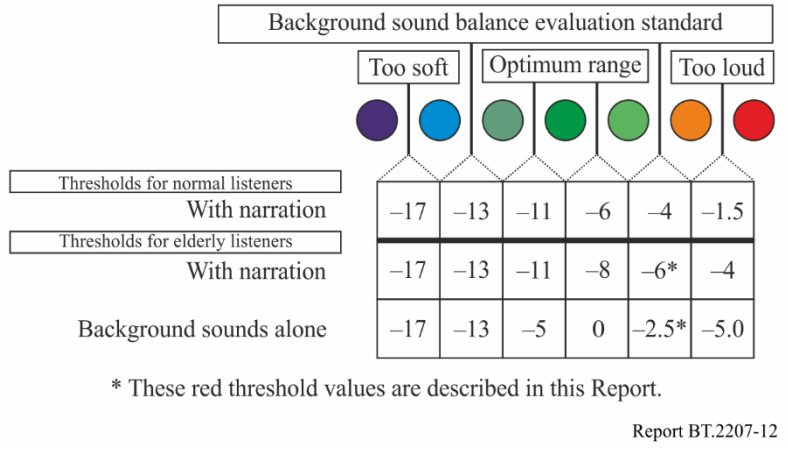


FIGURE 12

Loudness level thresholds when listening loudness level is 70 phon



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).

6 Easy-to-read language broadcasting service and language conversion support technology

The difficulty level of language used in broadcasts in general is based on the assumption that the audience consists of native speakers who are able-bodied listeners of a certain age. This means that many are likely to find this level of language rather difficult to understand – including non-native speakers, children, people with developmental impairments, and people with cognitive losses due to aging – which raises the prospect of a broadcast language barrier.

One way to solve this problem would be to provide broadcasting services tailored to the language comprehension level of these different groups. NHK has begun researching an easy Japanese broadcasting service tailored to the language proficiency level of foreign residents who are capable of speaking everyday conversational Japanese [1].

Japan currently has approximately 2.38 million non-native Japanese speakers, 1.9% of Japan's population – this population is so diverse that it is virtually impossible to provide broadcasts in all of the native languages of residents in Japan. At the same time, surveys have shown there is demand for broadcasts in easy Japanese [2], so there is certainly a need for such services.

The first objective was to come up with a service for converting the regular news content that is posted on the Web into easy Japanese content.

This involves three basic tasks: (1) figuring out the proper level of language difficulty that is tailored to the comprehension level of foreign residents who speak Japanese as a second language, (2) providing Web functions that further aid understanding of easy Japanese, and (3) developing a support system that translates ordinary Japanese into easy Japanese. Although these issues are still being worked out, a public trial service called NEWS WEB EASY was launched in April 2012 based on machine-aided human translations of daily news articles into easy Japanese. This report will propose solutions to the issues outlined above and go into a bit more detail about the trial NEWS WEB EASY service. At the end of the report, a brief summary and a number of future challenges will be presented.

6.1 What does easy Japanese mean?

The audience for these services is foreign residents learning Japanese as a second language who are already fairly fluent in conversational Japanese, but now want to learn how to read news articles and the newspaper. In other words, the focus is on foreign residents who have achieved pre-intermediate level Japanese. One way to define Japanese that matches this level of comprehension is to examine the standard process of Japanese language learning as a second language, and this suggests the following guidelines.

Vocabulary

Rewrite news articles by sticking as closely as possible to the elementary vocabulary that foreign residents learn at the initial stage of their study. For this purpose, the basic words listed in the test guidelines of the "Japanese language proficiency test" are currently being used [3]. Note that a number of terms not from this list are also used, including technical terms, proper names, and terms that frequently appear in news articles yet are difficult to reword in easy words.

Grammar

In terms of grammar, guidelines determine which functional expressions (such as passive, causative, hearsay, and intent expressions) are considered acceptable and define the degree of syntactic complexity of target phrases. Again, as with the vocabulary items, the policy was to use functional expressions that foreign residents are likely to learn at the beginning of their study. For example, the passive form was avoided as much as possible, since this is hard for people at a beginning level of Japanese to understand. However, there are a few exceptions. For example, considering the way news is quite often reported as hearsay, a few advanced level hearsay type expressions such as "to-shite-imasu (it has been reported that)" and "to-iukoto-desu (it is explained that)" were included. While these are generally not used in everyday conversation and are somewhat difficult, they are essential if one is to read and understand the news. Regarding syntactic complexity, the following guideline was adopted. Longer sentences typically tend to be more syntactically complex and harder to understand. To reduce syntactic complexity and facilitate clear understanding, long wordy sentences were broken into smaller sentences. Sentences in news copy that exceed 100 characters in length are extremely difficult to understand, so sentences are shortened to 60 characters or less.

Curtail information

There tends to be a lot of redundancy in Web news content. This is because news copy is often written based on a script prepared to be broadcast by voice over the radio. To streamline and simplify the Japanese, this duplication was eliminated. Supplemental information was also reduced as much as possible, which makes the easy Japanese version significantly shorter than the original news story.

6.2 Web service functions

Besides measures for simplifying Japanese itself as outlined above, a number of Web functions that aid the user in understanding the news were also provided. The following functions have been implemented in the public trial version of NEWS WEB EASY that is now available.

Furigana (ruby) characters

Japanese is written using a combination of Chinese characters (kanji), two types of Japanese phonetic symbols (hiragana and katakana), as well as alphabetical characters (romaji) and numbers. Kanji are notoriously difficult for foreigners to master because there are so many of them and because the same characters can be read in different ways depending on the context. Foreign residents thus often find themselves unable to understand the meaning of words written in kanji. To assist them in reading kanji, very small kana characters called furigana are printed above all kanji, indicating how they are correctly pronounced. This increases the chances of foreign readers being able to understand the meaning of kanji terms even if they are unable to read the kanji.

Glossaries

The basic approach is to write easy Japanese using elementary vocabulary, but it is generally not possible to convert all of the difficult terms to simple vocabulary. One solution might be to add a phrase explaining such terms in a sentence, but this would only lengthen the sentence and make it harder to understand. For this situation, glossaries are employed to explain difficult terminology. A glossary entry is accessed by merely positioning the cursor over the word on the NEWS WEB EASY trial screen. A popup is then displayed which explains the term. For the purposes of this trial, a dictionary for Japanese elementary school students was used to provide the glossary entries.

Proper nouns

Proper nouns – names of people, places, organizations, and so on – are unavoidable in news articles, and of course proper nouns are not included in any pre-existing glossary. Different classes of proper nouns are highlighted in different colours to draw the readers' attention. The reader may not know exactly what the terms mean, but at least he or she is able to differentiate the names of people, places, and organizations.

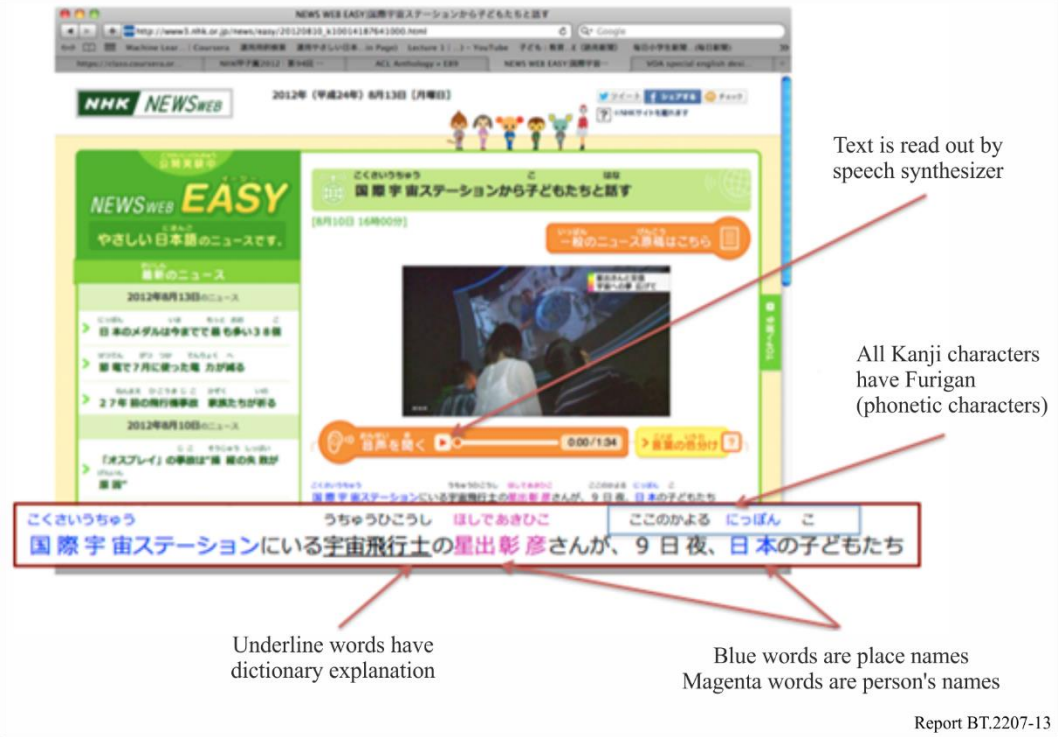
Text-to-speech

Some foreign residents have difficulty reading Japanese, yet are perfectly capable of understanding the text if it is read to them. NEWS WEB EASY features a text-to-synthesized voice function for people who fall into this category.

Figure 13 shows a screen shot of the NEWS WEB EASY trial, highlighting the features described.

FIGURE 13

Screen shot of the NEWS WEB EASY application



6.3 Rewrite support system

Translating articles into Easy Japanese is done in pairs by a news editor (a reporter) and a rewriter (a Japanese instructor trained in easy Japanese guidelines). The rewriter rewords the news stories without deleting any information, while the news editor deletes redundant and supplemental information and checks the accuracy of the rewritten content. The editors and rewriters are experts in their respective fields, but they do not have any experience in translating ordinary Japanese into easy Japanese. Rewriting work is far harder than one might imagine. To assist with the rewrite work, a special editor [4] and a model phrase search system have been developed.

Rewrite support editor

One of the challenges facing the rewriter and the editor is figuring out which parts of the text should be rewritten. It was noted earlier that difficult words (words that go beyond the elementary vocabulary level) and overly long phrases and sentences should be rewritten, but it is difficult to manually identify all of the places that should be reworded. To simplify this task, a function that automatically highlights these places and assigns a numerical index of difficulty was developed. Editors and rewriters are less likely to overlook spots marked for rewriting because they can see them at a glance. Another challenge is assessing the actual effects of rewriting. For example, it is sometimes difficult to determine whether rewriting actually simplifies the text or note specially if the editor is not well versed in the rules of easy Japanese. Here again, assistance is offered in the form of a function that assesses the overall difficulty of an article and then assigns a numerical value to the difficulty. The value factors in several criteria: the number of difficult words included in the article, the average length of sentences in the article, and the overall length of the article. The smaller the value, the simpler the article. By simply checking whether the value has gone up or down after rewriting, the editor and the rewriter know whether their efforts have improved the article or made it worse.

Model phrase search system

The rewrite support editor quickly identifies wordy or difficult passages, but does not offer alternative suggestions that might improve the article. For this, a database system was developed that continually stores articles before and after rewriting. The system automatically searches for and proposes simpler phrases to substitute for difficult phrases input by the rewriter. Using the system, for example, one can easily search for an alternative phrase to substitute for the common yet difficult hearsay expression "to-shite-imasu (it has been reported that)". This system makes it easy for rewriters to share experience and knowledge in the course of their work.

6.4 Future challenges

The NEWS WEB EASY trial has gone very smoothly since it was rolled out in April 2012, with two or three new articles released every day. Right now, news in easy Japanese is being assessed by polling users and by testing the comprehension of foreign residents using this service. Preliminary results suggest that the service has great promise and should be highly effective for conveying news to people at a pre-intermediate level of Japanese. Building on the promising results achieved so far, the easy Japanese guidelines will be fine-tuned to produce news articles that are even more transparent. An automatic translation function is also being developed that will significantly enhance the rewrite efficiency of the system.

References

[1] TANAKA, H. and MINO, H. [2010], Manual Translation Experiment of Broadcast News in Simple Japanese, Information Processing Society of Japan, IPSJ SIG Notes 2010-NL-199 (11), pp. 1-8, 11 Nov. 2010, (in Japanese).

[2] YONEKURA, R. [2012], Media Use and Information Behaviour of Foreign Residents in Japan When Disasters Occur, The NHK Monthly Report on Broadcast Research, Aug. 2012, (in Japanese).

[3] JAPAN FOUNDATION [2002], Association of International Education: Japanese Language Proficiency Test Content Specification, (in Japanese).

[4] MINO, H. and TANAKA, H. [To appear 2012], Rewrite Support Tool for Converting New Stories into Easy Japanese News: Targeting Foreigners Residing in Japan, Annual Conference of the Institute of Image Information and Television Engineers, (in Japanese).

7 Sound level adjustment system with a speech rate conversion (SRC) for channel-based stereo signals

As the average age of the television audience increases, broadcasters receive more comments from viewers that the narration in broadcast programmes is difficult to hear. There are two main reasons for that: (1) the background sound (music and sound effects) in broadcast programmes disturbs the viewers' recognition of narration, speech, comments, or dialogue; (2) the narration or speech is too rapid for an aging population. Concerning the issue of background sound in broadcast programmes, Nakamura et al. [1] reported reducing the background levels by only 3-6 dB lower than usual would make the narration easier to understand by a wider age range. Imai et al. [2] have shown that decreasing the speech rate when there is no background sound (as in a newscast) makes it easier to hear (see § 1). This section describes a prototype of a sound level adjustment system with a speech rate conversion for channel-based stereo signals.

7.1 Prototype speech rate conversion system

A block diagram of the prototype system is shown in Fig. 14. The stereo sound correlation (SSC) block divides the input stereo signals into an estimated speech signal and two estimated background sound signals. The spectral contrast enhancement (SCE) block enhances only the estimated speech. The speech rate conversion (SRC) block converts the speech rates of all the sound signals (estimated speech, estimated background sound, enhanced speech, and original sound). Using the speech/nonspeech segment information estimated by the voice activity detector, the system controls the speech rate and the gain of each signal.

The sound level adjustment system can adjust a mixing balance of the speech and background sound by varying the multipliers α, β, γ and η. The multipliers α and β adjust the reproduction level of the enhanced speech and the estimated speech, respectively. The multiplier γ adjusts the estimated background sound signal. The multiplier η adjusts the level of the original sound signal in the nonspeech segments. The background sound signal can be suppressed by decreasing γ and η. The gain control is performed depending on the speech/nonspeech segment information.

The SSC block employs an adaptive filter for background sound separation [3]. Using a time-varying filter, the SSC block can extract the correlated signal from the original stereo signal as the estimated speech signal. The uncorrelated signals, which are the estimated background sound signals, are obtained by subtracting the estimated speech signal from the original signals of the left and right channels.

The SCE block performs three procedures. In the first step, the estimated speech signal is divided into about 80 signals with different bandwidths. In the second step, the band signal levels with a peak in the spectral envelope are increased by filtering with steep characteristics, and in the third step, the band signal levels with a dip in the spectral envelope are suppressed by the filter of gentle curve characteristics. These combined filters enhance the contrast of the spectral envelope [4].

The SRC block adjusts the speech rate (see § 1).

The voice activity detector estimates the segmentations of the speech and nonspeech periods using amplitude information and the likelihood of phoneme recognition.

The prototype system (see Fig. 15) can replay video in synchronisation with the speech-rate-converted audio signal. This is realised by converting the frame rate of the video signal according to the SRC.

FIGURE 14

Block diagram of a sound level adjustment system with a speech rate conversion for stereo signals

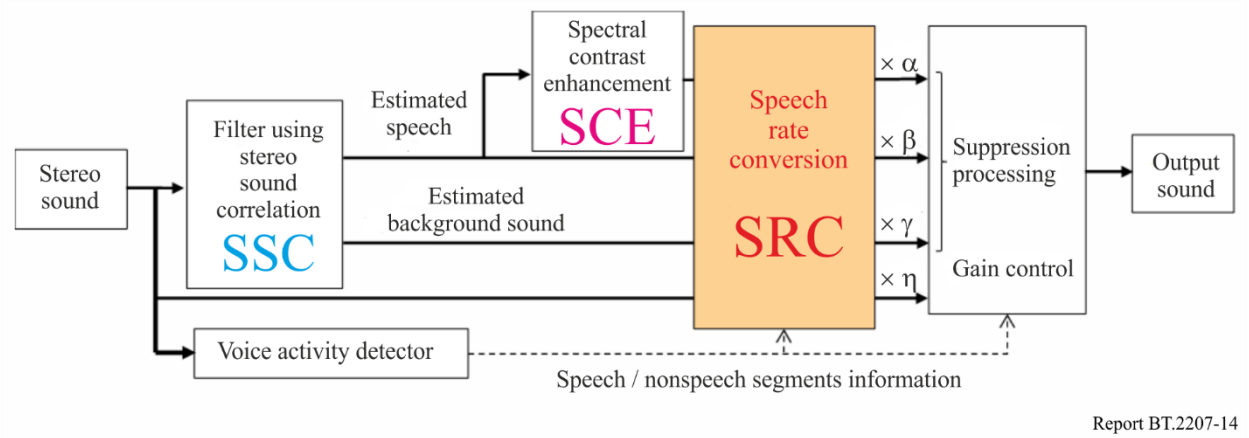
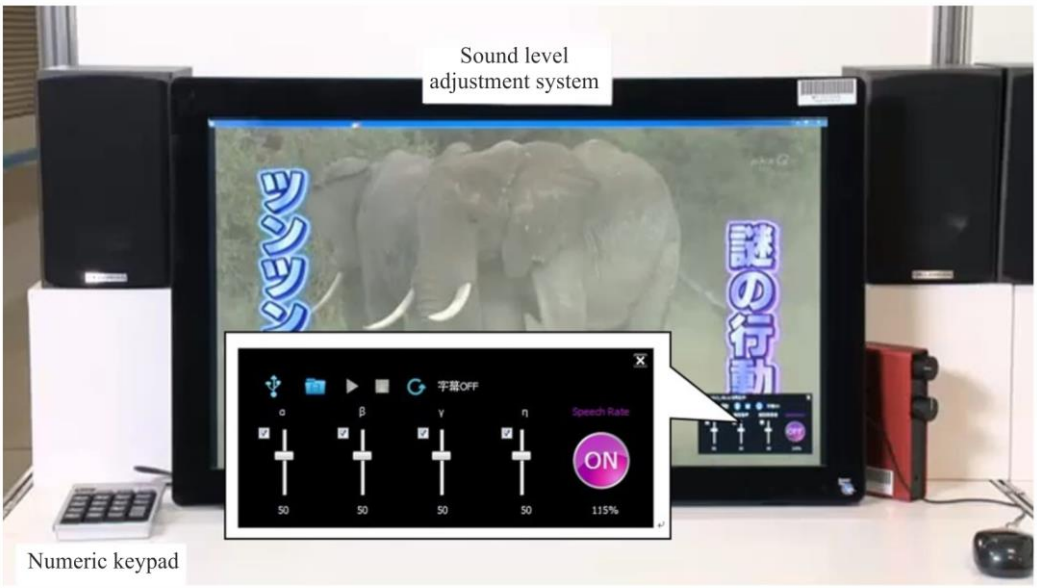


FIGURE 15

Prototype of the sound level adjustment system with a speech rate conversion for stereo signals



In cases without the SRC, an evaluation experiment showed that a 15 dB suppression ($\gamma = 0.178$) of the estimated background sound had the same effect of a 4.5 dB suppression of the actual background sound, whereas a 3-6 dB suppression is the target in accordance with [1]. The SCE with optimal parameters additionally improved the suppression effect of 1 or 2 dB [5], [6]. Therefore, the sound level adjustment system using both SSC and SCE achieved the target suppression of 6 dB on average through a 15 dB suppression of the estimated background sound.

7.2 Subjective evaluation

Subjective evaluation experiments were conducted to confirm how the prototype system could improve the audibility of stereo sound for an aging audience.

The effect of mixing balance adjustment was evaluated using selected test materials – each about 20 to 30 s long and including TV dramas – that had been reported to be difficult to hear. Using the prototype system, the mixing balance of the enhanced speech, the estimated speech, the estimated background sound was adjusted in the speech segment and the speech rate was changed 15% slower, i.e. a speech rate of about 6-7 mora/s, for older viewers prefer according to the results of a previous research [2]. Four sets of mixing balance were evaluated.

Sixteen subjects representative of an older audience (twelve subjects were normal hearing (NH) group and four subjects were lower hearing ability (LH) group) participated in the test conducted in a soundproof room. The LH group's hearing in 2-8 kHz frequency bands is particularly poor, indicating that high-frequency sounds are difficult to comprehend. The subjects were asked to judge the easiness of hearing using the seven-grade criteria to any pairs of four sets of mixing balance.

The result showed that when the mixing balance was adjusted to reduce the background sound, the average scores were higher than that for the original sound, demonstrating significant effects of background sound level adjustment [7].

Additionally, the effect of adding SRC to the sound level adjustment was also evaluated [7]. This result indicated that the sound level adjustment with SRC was more effective than the sound level adjustment without SRC for those with age related hearing loss.

References

[1] NAKAMURA, H., SAWAGUCHI, M., MASAOKA, K., WATANABE, K., YAMASAKI, Y., MIYASAKA, E., YASUOKA, M., SEKI, H. [2003] Better Audio Balance Broadcasting Service for Elderly People: Background Sound Levels of Television Programs for Easy Listening. Proc. Spring Meet. Acoust. Soc. Jpn., 1-5-5, pp. 455–456 (in Japanese).

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

[3] MURAYAMA, Y., HAMADA, H., KOMIYAMA, S., KAWABATA, Y. [2007] Adaptive Control for Advanced Reproduction of Narration Voice. 13th AES Regional Convention, Tokyo.

[4] TAKOU, R., SEIYAMA, N., IMAI, A., KOMORI, T., TAKAGI, T. [2012] A Study on Spectrum Contrast Enhancement for Sentence Speech Intelligibility in Noise for Elderly Persons. 2012 Proc. Autumn Meet. Acoust. Soc. Jpn., 2-Q-a8, pp. 531–532 (in Japanese).

[5] KOMORI, T., IMAI, A., SEIYAMA, N., TAKOU, R., TAKAGI, T., OIKAWA, Y. [2012] Development of a Broadcast Sound Receiver for Elderly Persons. Proc. 13th ICCHP, pp. 681–688, (Springer-Verlag, 2012).

[6] KOMORI, T., IMAI, A., SEIYAMA, N., TAKOU, R., TAKAGI, T., OIKAWA, Y. [2013] Development of Volume Balance Adjustment Device for Voices and Background Sounds Within Programs for Elderly People. 135th AES Convention Paper, 9010.

[7] KOMORI, T., IMAI, A., SEIYAMA, N., TAKOU, R., TAKAGI, T., OIKAWA, Y. [2014] Study of TV Sound Level Adjustment System for the Elderly with Speech Rate Conversion Function. 137th AES Convention Paper, 9167.