

DESCRIPTION

TRANSMISSION DEVICE, TRANSMISSION METHOD, RECEPTION DEVICE,
AND RECEPTION METHOD

5 TECHNICAL FIELD

[0001]

The present technology relates to a transmission device,
a transmission method, a reception device, and a reception
method, and particularly relates to a transmission device,
10 a transmission method, a reception device, and a reception
method that can enhance the accessibility for people with
visual disabilities.

BACKGROUND ART

15 [0002]

In the field of digital broadcasting, the accessibility
for people with visual disabilities is required (for example,
see Patent Document 1).

[0003]

20 Especially, the 21st Century Communications and Video
Accessibility Act of 2010 (CVAA), or the so-called American
accessibility act was enacted in the United States. Federal
Communications Commission (FCC) has published various
regulations on the accessibility based on the act.

25

CITATION LIST

PATENT DOCUMENT

[0004]

Patent Document 1: Japanese Patent Application Laid-Open No.

30 2009-204711

SUMMARY OF THE INVENTION

PROBLEMS TO BE SOLVED BY THE INVENTION

[0005]

By the way, when a user interface (UI), for example,
5 for program information is presented to people with visual
disabilities, the text information of the program information
is read aloud with a Text To Speech (TTS) engine. This is
a typical way to enhance the accessibility.

[0006]

10 However, when the TTS engine is used, the text
information is not necessarily read aloud as the producer of
the program information intends. It is uncertain whether
people with visual disabilities can receive the information
similar to the information that able-bodied people receive.
15 In light of the foregoing, a technology with which utterance
as the producer intends is surely performed so that people
with visual disabilities can receive the information similar
to the information that able-bodied people receive.

[0007]

20 In light of the foregoing, the present technology enables
the enhancement in the accessibility for people with visual
disabilities by surely performing utterance as the producer
intends.

25 SOLUTIONS TO PROBLEMS

[0008]

A transmission device according to a first aspect of
the present technology includes: a metadata generation unit
that generates metadata about text-to-speech utterance of
30 display information, the text-to-speech utterance being
performed as a producer intends; an electronic program

information generation unit that generates electronic program information including the metadata; and a transmission unit that transmits the electronic program information to a reception device capable of displaying the display information.

[0009]

The metadata can include information about utterance of a character string of which pronunciation is not uniquely determined, or a character string difficult to pronounce.

The display information can include information or an icon of a content.

[0011]

A content acquisition unit that acquires the content can be further included, and the transmission unit can transmit the electronic program information together with the content using a digital broadcast signal.

[0012]

The electronic program information can comply with Electronic Service Guide (ESG) prescribed in Open Mobile Alliance-Mobile Broadcast Services Enabler Suite (OMA-BCAST), the metadata can be described in a Speech Synthesis Markup Language (SSML) format, and a predetermined fragment included in the ESG can include address information indicating an address at which a file of the metadata described in the SSML format is acquired, or content of the metadata described in the SSML format.

[0013]

The transmission device can be an independent device, or can be an internal block included in a device.

[0014]

The transmission method according to the first aspect of the present technology corresponds to the transmission device according to the first aspect of the present technology.
[0015]

5 In the transmission device and transmission method according to the first aspect of the present technology, the metadata about the text-to-speech utterance that the producer intends is generated for display information, and the electronic program information including the metadata is
10 generated. Then, the electronic program information is transmitted to a reception device capable of displaying the display information.
[0016]

A reception device according to a second aspect of the
15 present technology includes: a reception unit that receives electronic program information, the electronic program information being transmitted from a transmission device, the electronic program information including the metadata about text-to-speech utterance of display information, the
20 text-to-speech utterance being performed as a producer intends; a metadata acquisition unit that acquires the metadata included in the electronic program information; and a text-to-speech reading aloud unit that reads the display information aloud on the basis of the metadata.
25 [0017]

The metadata can include information about utterance of a character string of which pronunciation is not uniquely determined, or a character string difficult to pronounce.
[0018]

30 The display information can include information or an icon of a content.

[0019]

The reception unit can receive the electronic program information transmitted together with the content as a digital broadcast signal.

5 [0020]

The electronic program information can comply with ESG prescribed in OMA-BCAST, the metadata can be described in an SSML format, a predetermined fragment included in the ESG can include address information indicating an address at which
10 a file of the metadata described in the SSML format is acquired, or content of the metadata described in the SSML format, and the metadata acquisition unit can acquire the file of the metadata in accordance with the address information, or acquires the metadata from the fragment.

15 [0021]

The reception device can be an independent device, or can be an internal block included in a device.

[0022]

The reception method according to the second aspect of
20 the present technology corresponds to the reception device according to the second aspect of the present technology.

[0023]

In the reception device and reception method according to the second aspect of the present technology, the electronic
25 program information, which includes the metadata about the text-to-speech utterance of the display information as the producer intends and is transmitted from the transmission device, is received. Then, the metadata included in the electronic program information is acquired, and the display
30 information is read aloud on the basis of the metadata.

EFFECTS OF THE INVENTION

[0024]

According to the first aspect and the second aspect of the present technology, the accessibility for people with visual disabilities can be enhanced.

[0025]

Note that the effects of the present technology are not necessarily limited to the effect described above, and can be any one of the effects described herein.

10

BRIEF DESCRIPTION OF DRAWINGS

[0026]

Fig. 1 is a diagram of an example in which program information and a title are read aloud.

Fig. 2 is a diagram of an example in which an icon is read aloud.

Fig. 3 is an explanatory diagram of an example in which text information is read aloud with an existing TTS engine.

Fig. 4 is an explanatory diagram of an example in which text information is read aloud with an existing TTS engine.

Fig. 5 is an explanatory diagram of an example in which text information is read aloud with a TTS engine to which the present technology is applied.

Fig. 6 is an explanatory diagram of an example in which text information is read aloud with a TTS engine to which the present technology is applied.

Fig. 7 is a diagram of an exemplary configuration of a broadcasting system to which the present technology is applied.

Fig. 8 is a diagram of an exemplary configuration of a transmission device to which the present technology is

applied.

Fig. 9 is a diagram of an exemplary configuration of a reception device to which the present technology is applied.

Fig. 10 is a diagram of an exemplary configuration of
5 ESG.

Fig. 11 is a diagram of an exemplary configuration of a service fragment of ESG.

Fig. 12 is a diagram of an exemplary configuration of a content fragment of ESG.

10 Fig. 13 is a diagram of an exemplary configuration of extended ESG.

Fig. 14 is a diagram of another exemplary configuration of extended ESG.

15 Fig. 15 is a diagram of the detailed configuration of a PhoneticInfoURI element.

Fig. 16 is a diagram of the detailed configuration of a PhoneticInfo element.

Fig. 17 is a diagram of an exemplary description of a sub element in an SSML format.

20 Fig. 18 is a diagram of an exemplary description of a phoneme element in an SSML format.

Fig. 19 is a diagram of an exemplary description of an audio element in an SSML format.

25 Fig. 20 is an explanatory flowchart of a transmitting process.

Fig. 21 is an explanatory flowchart of a receiving process.

Fig. 22 is a diagram of an exemplary configuration of a computer.

30

MODE FOR CARRYING OUT THE INVENTION

[0027]

The embodiments of the present technology will be described hereinafter with reference to the appended drawings. Note that the embodiments will be described in the following
5 order.

[0028]

1. Outline of Text-to-speech Utterance Metadata of Present Technology
2. Configuration of System
- 10 3. Arrangement of Text-to-speech Utterance Metadata with Extension of ESG
4. Exemplary Description of Text-to-speech Utterance Metadata
5. Flow of Process Performed in Each Device
- 15 6. Exemplary Variation
7. Configuration of Computer

[0029]

<1. Outline of Text-to-speech Utterance Metadata of Present Technology>

20 [0030]

Among the Federal communications commission's regulations related to the American accessibility act, the regulation on a user interface (FCC Report & Order (FCC 13-138) released October 31, 2013. C.F.R. Title 47 §79.107, 108)
25 requires that the user interfaces on a receiver capable of displaying a video program, such as a television receiver, be accessible for people with visual disabilities.

[0031]

Specifically, when a screen of Electronic Service Guide
30 (ESG), namely, an ESG screen is displayed as illustrated in Fig. 1, for example, the program information and a title are

read aloud. This can provide the information necessary to select a station of a broadcast program to people with visual disabilities.

[0032]

5 Alternatively, when a menu screen is displayed as illustrated in Fig. 2, for example, the content of a service indicated by each icon is read aloud. This can provide the content of the service to people with visual disabilities.

[0033]

10 As described above, the information about the user interface displayed on the receiver is read aloud. This enables people with visual disabilities to acquire the information about the user interface and perform various operations of the receiver.

15 [0034]

 Meanwhile, the channel information or program information for selecting the station of a broadcast program (program) is provided as the ESG information from a transmitter, for example, of a broadcast station to receivers. The ESG
20 information mainly includes text information or logo data. Then, the receiver generates and displays an ESG screen for selecting the station of a broadcast program on the basis of the ESG information.

[0035]

25 As described above, when an ESG screen is displayed, the accessibility of the user interface for people with visual disabilities is required and, for example, a title or program information is read aloud. In such a case, the text information of the title or the program information is typically read aloud
30 with a Text To Speech (TTS) engine. The TTS engine is a voice synthesizer (Text To Speech Synthesizer) that can artificially

create human voice from text information.

[0036]

However, the TTS engine does not necessarily read aloud the text information as the producer of the user interface intends. It is uncertain that people with visual disabilities can receive the information similar to the information that able-bodied people receive.

[0037]

Specifically, for example, when the text information is "AAA" as illustrated in Fig. 3, the text information can be said either "triple A" or "A A A". The pronunciation is not uniquely determined. Thus, it is difficult for the TTS engine to determine how to read the text information aloud. As a result, the text information is likely not to be read aloud as the producer intends.

[0038]

Alternatively, for example, when the text information is "Caius College" as illustrated in Fig. 4, the text information is a proper noun difficult to pronounce. Thus, it is difficult for the TTS engine to determine how to read the text information aloud. As a result, the text information is likely not to be read aloud as the producer intends.

[0039]

When the pronunciation of the text information is not uniquely determined, or when the text information is, for example, a proper noun difficult to pronounce as described above, there is possibility that the text information is not read aloud as the producer intends. Thus, there is need for a technique with which utterance as the produce intends is surely performed and people with visual disabilities can receive the information similar to the information that

able-bodied people receive.

[0040]

In light of the foregoing, according to the present technology, the information about text-to-speech utterance that the producer intends (hereinafter, referred to as "text-to-speech utterance metadata") is provided to the TTS engine so that the TTS engine can emit the voice that the producer intends in order to perform text-to-speech utterance as the producer intends to read the display information, for example, of the user interface aloud. Note that the text-to-speech utterance metadata can be included in the ESG information.

[0041]

Specifically, as illustrated in Fig. 5, for example, when the text information is "AAA", the "triple A" indicating the pronunciation of the text information is provided as the text-to-speech utterance metadata to the TTS engine. This enables the TTS engine to read the text information aloud as "triple A" on the basis of the text-to-speech utterance metadata.

[0042]

In other words, when the text information "AAA" is input, the TTS engine in Fig. 3 does not determine which pronunciation "triple A" or "A A A" is correct. On the other hand, the text-to-speech utterance metadata "triple A" is input to the TTS engine in Fig. 5, and the TTS engine can read the "triple A" aloud in accordance with the text-to-speech utterance metadata. Thus, the text-to-speech utterance as the producer intends is performed.

[0043]

Alternatively, for example, when the text information is "Caius College", the phoneme information of the text

information is provided as the text-to-speech utterance metadata to the TTS engine as illustrated in Fig. 6. This enables the TTS engine to read the text information aloud as "keys college" on the basis of the text-to-speech utterance metadata.

[0044]

In other words, when the text information "Caius College" is input, the TTS engine in Fig. 4 does not determine how to pronounce the text information because the text information is a proper noun difficult to pronounce. On the other hand, the phoneme information is input as the text-to-speech utterance metadata to the TTS engine in Fig. 6, and the TTS engine can read the text information aloud as "keys college" in accordance with the text-to-speech utterance metadata. Thus, the text-to-speech utterance as the producer intends is performed.

[0045]

Providing the text-to-speech utterance metadata to the TTS engine as described above allows for surely reading the text information aloud as the producer intends even when, for example, the pronunciation of the text information is not uniquely determined, or the text information is a proper noun difficult to pronounce. This enables people with visual disabilities to receive the information similar to the information able-bodied people receive.

[0046]

<2. Configuration of System>

[0047]

(Exemplary Configuration of Broadcasting System)

Fig. 7 is a diagram of an exemplary configuration of a broadcasting system to which the present technology is

applied.

[0048]

A broadcasting system 1 provides the content, for example, of a broadcast program, and can make the display information, for example, of a user interface accessible for people with visual disabilities. The broadcasting system 1 includes a transmission device 10 and a reception device 20.

[0049]

The transmission device 10 is managed, for example, by a broadcast station that provides a terrestrial digital broadcasting service. The transmission device 10 transmits the content, for example, of a broadcast program with a digital broadcast signal. Furthermore, the transmission device 10 generates ESG information including text-to-speech utterance metadata and puts the ESG information in the digital broadcast signal and transmits the digital broadcast signal.

[0050]

The reception device 20 includes, for example, a television receiver or a set-top box and is installed, for example, at the house of each user. The reception device 20 receives the digital broadcast signal transmitted from the transmission device 10, and outputs the video and voice of the content, for example, of a broadcast program.

[0051]

Furthermore, the reception device 20 includes a TTS engine so as to read the display information, for example, of a user interface aloud on the basis of the text-to-speech utterance metadata included in the ESG information when the display information of the user interface is displayed.

[0052]

This configuration causes the TTS engine to read the

display information such as text information aloud in accordance with the text-to-speech utterance metadata. Thus, the display information is read aloud surely as the producer intends even when, for example, the pronunciation of the text information is not uniquely determined, or the text information is a proper noun difficult to pronounce.

[0053]

Note that, the broadcasting system 1 in Fig. 7 includes a transmission device 10. However, a plurality of transmission devices 10 is installed at a plurality of broadcast stations, respectively. Similarly, the broadcasting system 1 in Fig. 7 includes a reception device 20. However, a plurality of reception devices 20 is installed at a plurality of houses of users, respectively.

[0054]

(Exemplary Configuration of Transmission Device)

Fig. 8 is a diagram of an exemplary configuration of the transmission device of Fig. 7.

[0055]

Fig. 8 illustrates that the transmission device 10 includes a content acquisition unit 111, a text-to-speech utterance metadata generation unit 112, an ESG information generation unit 113, a stream generation unit 114, and a transmission unit 115.

[0056]

The content acquisition unit 111 acquires the content, for example, of a broadcast program and provides the content to the stream generation unit 114. Furthermore, the content acquisition unit 111 can process the contents, for example, in a process for encoding the content or a process for converting the format.

[0057]

Note that the content is, for example, acquired from the storage of recorded contents in accordance with the broadcasting time of the contents, or acquired from the studio or a location of live broadcasting.

[0058]

The text-to-speech utterance metadata generation unit 112 generates text-to-speech utterance metadata, for example, in response to the instructions from the producer of the user interface, and provides the text-to-speech utterance metadata to the ESG information generation unit 113. Note that, for example, the information indicating the pronunciation of the text information when the pronunciation is not uniquely determined, or the phoneme information when the text information is a proper noun difficult to pronounce is generated as the text-to-speech utterance metadata.

[0059]

In this example, there are two types of text-to-speech utterance metadata stored in the ESG information. One is the data describing the address information to acquire the text-to-speech utterance metadata, and the other is the data describing the content of the text-to-speech utterance metadata. When the address information is described in the text-to-speech utterance metadata, the content of the text-to-speech utterance metadata is described in a file acquired in accordance with the address information (hereinafter, referred to as a "text-to-speech utterance metadata file").

[0060]

In other words, when the text-to-speech utterance metadata generation unit 112 generates the text-to-speech

utterance metadata including the address information and provides the text-to-speech utterance metadata to the ESG information generation unit 113, the text-to-speech utterance metadata generation unit 112 generates a text-to-speech utterance metadata file to be acquired in accordance with the address information and provides the text-to-speech utterance metadata file to the stream generation unit 114. On the other hand, when the text-to-speech utterance metadata includes the content of the text-to-speech utterance metadata, the text-to-speech utterance metadata generation unit 112 does not need to generate a text-to-speech utterance metadata file, and thus only provides the text-to-speech utterance metadata to the ESG information generation unit 113.

[0061]

The ESG information generation unit 113 generates ESG information as channel information for selecting the station of the content, for example, of a broadcast program. Furthermore, the ESG information generation unit 113 stores (places) the text-to-speech utterance metadata provided from the text-to-speech utterance metadata generation unit 112 in the ESG information. The ESG information generation unit 113 provides the ESG information including the text-to-speech utterance metadata to the stream generation unit 114.

[0062]

The stream generation unit 114 generates a stream in compliance with a predetermined standard on the basis of the content data provided from the content acquisition unit 111 and the ESG information provided from the ESG information generation unit 113, and provides the stream to the transmission unit 115.

[0063]

Alternatively, when the text-to-speech utterance metadata included in the ESG information provided from the ESG information generation unit 113 includes the address information, the text-to-speech utterance metadata file is provided to the stream generation unit 114 from the text-to-speech utterance metadata generation unit 112. In this case, the stream generation unit 114 generates a stream in compliance with a predetermined standard on the basis of the content data provided from the content acquisition unit 111, the text-to-speech utterance metadata file provided from the text-to-speech utterance metadata generation unit 112, and the ESG information provided from the ESG information generation unit 113.

[0064]

The transmission unit 115 processes the stream provided from the stream generation unit 114, for example, in a digital modulation process, and transmits the stream as a digital broadcast signal through an antenna 116.

[0065]

Note that all of the functional blocks in the transmission device 10 of Fig. 8 are not necessarily arranged in a device. At least some of the functional blocks can be arranged as a device independent from the other functional blocks. For example, the text-to-speech utterance metadata generation unit 112 or the ESG information generation unit 113 can be provided as a function of a server on the Internet. In the example, the transmission device 10 acquires and processes the text-to-speech utterance metadata or ESG information provided from the server.

[0066]

(Exemplary Configuration of Reception Device)

Fig. 9 is a diagram of an exemplary configuration of the reception device of Fig. 7.

[0067]

Fig. 9 illustrates that the reception device 20 includes
5 a reception unit 212, a stream separation unit 213, a
reproduction unit 214, a display unit 215, a loudspeaker 216,
an ESG information acquisition unit 217, a text-to-speech
utterance metadata acquisition unit 218, and a TTS engine 219.

[0068]

10 The reception unit 212 processes the digital broadcast
signal received with the antenna 211, for example, in a decoding
process, and provides the stream generated by the process to
the stream separation unit 213.

[0069]

15 The stream separation unit 213 separates the content
data and the ESG information from the stream provided from
the reception unit 212, and provides the content data to the
reproduction unit 214, and the ESG information to the ESG
information acquisition unit 217.

20 [0070]

The reproduction unit 214 displays the picture of the
content on the display unit 215 on the basis of the content
data provided from the stream separation unit 213, and outputs
the voice of the content from the loudspeaker 216. This
25 reproduces the content, for example, of a broadcast program.

[0071]

The ESG information acquisition unit 217 acquires the
ESG information provided from the stream separation unit 213.
For example, when the user instructs the reception device 20
30 to display an ESG screen, the ESG information acquisition unit
217 provides the ESG information to the reproduction unit 214.

The reproduction unit 214 generates an ESG screen on the basis of the ESG information provided from the ESG information acquisition unit 217, and displays the ESG screen on the display unit 215.

5 [0072]

Furthermore, the ESG information acquisition unit 217 provides the text-to-speech utterance metadata included in the ESG information to the text-to-speech utterance metadata acquisition unit 218. The text-to-speech utterance metadata acquisition unit 218 acquires the text-to-speech utterance metadata provided from the ESG information acquisition unit 217.

[0073]

There are two types of text-to-speech utterance metadata, the data describing the address information to acquire the text-to-speech utterance metadata, and the data describing the content of the text-to-speech utterance metadata in this example as described above.

[0074]

20 In other words, when the text-to-speech utterance metadata includes the address information, the text-to-speech utterance metadata acquisition unit 218 acquires the text-to-speech utterance metadata file from the stream separated by the stream separation unit 213 on the basis of the included address information, and provides the text-to-speech utterance metadata including the content provided from the acquired text-to-speech utterance metadata file to the TTS engine 219. On the other hand, when the text-to-speech utterance metadata includes the content, the text-to-speech utterance metadata acquisition unit 218 provides the text-to-speech utterance metadata without any

process to the TTS engine 219.

[0075]

The TTS engine 219 reads the display information, for example, of a user interface aloud on the basis of the text-to-speech utterance metadata provided from the text-to-speech utterance metadata acquisition unit 218, and outputs the voice from the loudspeaker 216.

[0076]

For example, an ESG screen is displayed on the display unit 215, and a title or program information is to be read aloud so that the text information is accessible for people with visual disabilities. However, the pronunciation of the text information is not uniquely determined. In such an example, the TTS engine 219 enables the text information to be read aloud as the producer intends in accordance with the text-to-speech utterance metadata. This enables people with visual disabilities to receive the information similar to the information that the able-bodied people receive.

[0077]

Note that Fig. 9 illustrates that the display unit 215 and the loudspeaker 216 are internally provided in the reception device 20. However, the display unit 215 and the loudspeaker 216 can be provided as another external device.

[0078]

<3. Arrangement of Text-to-speech Utterance Metadata with Extension of ESG>

[0079]

Next, the ESG information in which text-to-speech utterance metadata is stored will be described in detail. Note that the specifications of the Electronic Service Guide (ESG) are designed by Open Mobile Alliance (OMA) that is an

organization for designing the standards of mobile phones.
The ESG information in which the text-to-speech utterance
metadata is stored is also in compliance with the ESG prescribed
in the OMA-Mobile Broadcast Services Enabler Suite

5 (OMA-BCAST).

[0080]

(Configuration of ESG)

Fig. 10 is a diagram of an exemplary configuration of
ESG. Note that each of the lines connecting fragments in Fig.
10 10 shows the cross-reference between the connected fragments.

[0081]

Fig. 10 illustrates that the ESG includes fragments
having different purposes, and is divided into four groups
of Administrative, Provisioning, Core, and Access in
15 accordance with the purpose of the fragment.

[0082]

The Administrative is a group that provides basic
information for receiving the ESG information. The group of
Administrative includes a "ServiceGuideDeliveryDescriptor".
20 The "ServiceGuideDeliveryDescriptor" provides the
information about the channel on which a plurality of service
guide fragments can be received, the scheduling information
about the channel, and the update information to the reception
device 20. This enables the reception device 20 to receive
25 only necessary ESG information at a proper time.

[0083]

The Provisioning is a group that provides the fee
information about the service reception. The group of
Provisioning includes a Purchase Item, Purchase Data, and a
30 Purchase Channel. The Purchase Item provides the fee
information about a service or a bundle of services. The

Purchase Data provides the information indicating which way the user can pay the fee through. The Purchase Channel provides the information about the system in which the user can actually purchase a service.

5 [0084]

Note that the text-to-speech utterance metadata or the address information indicating the address at which the text-to-speech utterance metadata is acquired can be stored in the fragment of each of the Purchase Item, the Purchase Data, and the Purchase Channel. A method of storing the text-to-speech utterance metadata in the fragment of each of the Purchase Item, the Purchase Data, and the Purchase Channel will be described below with reference to Fig. 13.

[0085]

15 The Core is a group that provides the information about a service. The group of Core includes Service, Schedule, and Content. The service provides metadata including the content of a channel service and the control information related to the content of the channel service. The Schedule provides the metadata including the schedule of the delivery of the content, and the control information related to the schedule of the delivery of the content. The Content provides the metadata including the content of the service and the control information related to the content of the service.

25 [0086]

Note that the text-to-speech utterance metadata or the address information indicating the address at which the text-to-speech utterance metadata is acquired can be stored in each fragment of the Service and the Content. Fig. 11 illustrates an exemplary configuration of the Service Fragment. Fig. 12 illustrates an exemplary configuration of the Content

Fragment. A method for storing the text-to-speech utterance metadata in each fragment of the Service and the Content will be described below with reference to Fig. 13.

[0087]

5 The Access is a group that provides the service access information indicating a method for receiving the service of the group of Core and the specific information about the session in which the content of the service is transmitted. The Access enables the reception device 20 to access a service. The group
10 of Access includes Access, and Session Description.

[0088]

 The Access in the group of Access provides a method for enabling the reception device 20 to access some additional services on the basis of a service by providing a plurality
15 of methods for accessing the service. The Session Description provides the session information about the service transmitted in the service access defined by an Access Fragment.

[0089]

 In addition to the four groups, the ESG includes Preview
20 Data, and Interactivity Data. The Preview Data provides a preview or an icon for a service and a content. The Interactivity Data provides the metadata about the application of the service and the content.

[0090]

25 Note that the text-to-speech utterance metadata or the address information indicating the address at which the text-to-speech utterance metadata is acquired can be stored in the Preview Data Fragment. A method for storing the text-to-speech utterance metadata in the Preview Data Fragment
30 will be described below with reference to Fig. 14.

[0091]

(Exemplary Configuration of Extended ESG)

Fig. 13 is a diagram of an exemplary configuration of ESG extended in order to store the text-to-speech utterance metadata or the address information indicating the address at which the text-to-speech utterance metadata is acquired. Note that Fig. 13 illustrates that the Service Fragment, the Content Fragment, the Purchase Item Fragment, the Purchase Data Fragment, and the Purchase Channel are extended among the fragments included in the ESG.

10 [0092]

In other words, the extended fragments include a Name element and a Description element. Thus, each of the fragment is extended so that a PhoneticInfoURI element or a PhoneticInfo element is added to the Name element and the Description element. Note that the PhoneticInfoURI element or the PhoneticInfo element can be added to the PrivateExt element of the extended fragment.

15

[0093]

Fig. 13 illustrates that the name of a content fragment is designated in the Name element. The Name element includes the PhoneticInfoURI element, the PhoneticInfo element, and a Type attribute as the child elements.

20

[0094]

The address information for acquiring the text-to-speech utterance metadata is designated in the PhoneticInfoURI element. The Type attribute is used together with the PhoneticInfoURI element as a pair. The type information indicating the type of text-to-speech utterance metadata is designated in the Type attribute.

25

30 [0095]

For example, a Uniform Resource Identifier (URI) is

designated as the address information. Alternatively, for example, when a text-to-speech utterance metadata file is transmitted in a File Delivery over Unidirectional Transport (FLUTE) session, the address information for acquiring the text-to-speech utterance metadata file transmitted in the FLUTE session is designated. Note that the text-to-speech utterance metadata can be described in Speech Synthesis Markup Language (SSML) that is a voice synthesis markup language. [0096]

10 The content of the text-to-speech utterance metadata is described in the PhoneticInfo element. For example, the content of the text-to-speech utterance metadata is described in SSML. The Type attribute is used together with the PhoneticInfo element as a pair. The type information indicating the type of text-to-speech utterance metadata is designated in the Type attribute. [0097]

 Note that, when "1 .. N" is designated as the number of occurrence (Cardinality) in Fig. 13, the element or attribute is designated once or more. When "0 .. N" is designated as the number of occurrence (Cardinality), it is arbitrarily determined whether the element or attribute is designated once or more. Alternatively, when "0 .. 1" is designated as the number of occurrence (Cardinality), it is arbitrarily determined whether the element or attribute is designated. [0098]

 Accordingly, the PhoneticInfoURI element, the PhoneticInfo element, and the Type attribute that are the child elements of the Name element are optional elements or attribute. Thus, not only either of the PhoneticInfoURI element or the

PhoneticInfo element is placed but also both of the PhoneticInfoURI element and the PhoneticInfo element can be placed.

[0099]

5 Furthermore, the Description element in Fig. 13 includes a PhoneticInfoURI element, a PhoneticInfo element, and a Type attribute as the child elements. In other words, the child elements of the Description element are similar to the child elements of the Name element.

10 [0100]

 Specifically, the address information for acquiring the text-to-speech utterance metadata is designated in the PhoneticInfoURI element. The Type attribute is used together with the PhoneticInfoURI element as a pair. The type
15 information indicating the type of text-to-speech utterance metadata is designated in the Type attribute. Furthermore, the content of the text-to-speech utterance metadata is described in the PhoneticInfo element. The Type attribute is used together with the PhoneticInfo element as a pair. The
20 type information indicating the type of text-to-speech utterance metadata is designated in the Type attribute.

[0101]

 Note that, as for the PhoneticInfoURI element and the PhoneticInfo element that are the child elements of the
25 Description element, either of the PhoneticInfoURI element or the PhoneticInfo element can be placed, or both of the PhoneticInfoURI element and the PhoneticInfo element can be placed.

[0102]

30 Fig. 14 is a diagram of another configuration of ESG extended in order to store the text-to-speech utterance

metadata or the address information indicating the address at which the text-to-speech utterance metadata is acquired. Note that Fig. 14 illustrates that the Preview Data Fragment is extended among the fragments included in the ESG.

5 [0103]

In other words, the Preview Data Fragment includes a Picture element, and thus the Preview Data Fragment is extended so that a PhoneticInfoURI element or a PhoneticInfo element is added to the relativePreference attribute of the Picture
10 element. Note that the PhoneticInfoURI element or the PhoneticInfo element can be added to the PrivateExt element of the Preview Data Fragment.

[0104]

Fig. 14 illustrates that a preview or an icon for a service
15 and a content is defined in the Picture element. The Picture element includes a PhoneticInfoURI element, a PhoneticInfo element, and a Type attribute as the child elements. In other words, the child elements of the Picture element are similar to the child elements of the Name element and the child elements
20 Description element described above.

[0105]

Specifically, the address information for acquiring the text-to-speech utterance metadata is designated in the PhoneticInfoURI element. The Type attribute is used together
25 with the PhoneticInfoURI element as a pair. The type information indicating the type of text-to-speech utterance metadata is designated in the Type attribute. Furthermore, the content of the text-to-speech utterance metadata is described in the PhoneticInfo element. The Type attribute
30 is used together with the PhoneticInfo element as a pair. The type information indicating the type of text-to-speech

utterance metadata is designated in the Type attribute.

[0106]

Note that, as for the PhoneticInfoURI element and the
PhoneticInfo element that are the child elements of the Picture
element, either of the PhoneticInfoURI element or the
PhoneticInfo element can be placed, or both of the
PhoneticInfoURI element and the PhoneticInfo element can be
placed.

[0107]

10 (Exemplary Configuration of PhoneticInfoURI Element)

Fig. 15 is a diagram of the detailed configuration of
a PhoneticInfoURI element in extended ESG.

[0108]

The PhoneticInfoURI element of Fig. 15 is described as
the child element of the Name element or Description element,
for example, of a Service Fragment, or as the child element
of the picture element of a Preview Data Fragment. The address
information for acquiring the text-to-speech utterance
metadata is designated in the PhoneticInfoURI element.

20 [0109]

Furthermore, an identification URI of the encoding
format of the text-to-speech utterance metadata is designated
as the type attribute of the PhoneticInfoURI element.

[0110]

25 For example, a text-to-speech utterance metadata file
is transmitted in the FLUTE session, and thus the address
information for acquiring the text-to-speech utterance
metadata file transmitted in the FLUTE session is designated
in the PhoneticInfoURI element.

30 [0111]

(Exemplary Configuration of PhoneticInfo Element)

Fig. 16 is a diagram of the detailed configuration of the PhoneticInfo element in extended ESG.

[0112]

Fig. 16 illustrates that the PhoneticInfo element is described as the child element of the Name element or Description element, for example, of a Service Fragment, or as the child element of the picture element of a Preview Data Fragment. The content of the text-to-speech utterance metadata is described in the PhoneticInfo element.

10 [0113]

Furthermore, an identification URI of the encoding format of the text-to-speech utterance metadata is designated as the type attribute of the PhoneticInfo element.

[0114]

15 For example, the content of the text-to-speech utterance metadata is described in SSML that is a speech synthesis markup language, and stored as the text information between the start tag and end tag of the PhoneticInfo element.

[0115]

20 Note that, at least one of the PhoneticInfoURI element and the PhoneticInfo element is described in the extended fragment in the ESG. Additionally, the text-to-speech utterance metadata is designated by the PhoneticInfoURI element or the PhoneticInfo element, and thus sometimes referred to as a "PhoneticInfo object".

25

[0116]

<4. Exemplary Description of Text-to-speech Utterance Metadata>

[0117]

30 As described above, the text-to-speech utterance metadata can be described, for example, in SSML that is a speech

synthesis markup language. The SSML is recommended by World Wide Web Consortium (W3C) with the aim of enabling usage of a higher-quality speech synthesizing function. Using SSML allows for close and appropriate control of the elements
 5 necessary for speech synthesis, such as pronunciation, volume, and rhythm. Hereinafter, exemplary description of a document in the SSML format will be described with reference to Figs. 17 to 19.

[0118]

10 (Sub Element)

Fig. 17 is a diagram of exemplary description of a sub element in the SSML format.

[0119]

The sub element is used to replace text information with
 15 another text information. The text information for text-to-speech utterance is designated in an alias attribute. For example, in Fig. 17, the text information "W3C" is replaced with the text information "World Wide Web Consortium" for text-to-speech utterance and read aloud.

20 [0120]

Using the sub element allows for designation of the information indicating the pronunciation of text information, for example, when the pronunciation of the text information is not uniquely determined.

25 [0121]

(Phoneme Element)

Fig. 18 is a diagram of exemplary description of a phoneme element in the SSML format.

[0122]

30 The phoneme element is used to give the pronunciation transcribed by phonemes/phonetic alphabet to the described

text information. An alphabet attribute and a ph attribute can be designated in the phoneme element. The phonetic characters of phonemes/phonetic alphabet are designated in the alphabet attribute. The character string of phonemes/phonemes/phonetic alphabet is designated in the ph attribute. For example, in Fig. 18, the pronunciation of the text information "La vita e bella" is designated in the ph attribute. Note that the "ipa" designated in the alphabet attribute indicates that the phonetic characters correspond to the reading symbols of International Phonetic Alphabet (IPA).

[0123]

Using the phoneme element allows for designation, for example, of the phoneme information of text information, for example, when the text information is a proper noun difficult to pronounce.

[0124]

(Audio Element)

Fig. 19 is a diagram of exemplary description of an audio element in the SSML format.

[0125]

The audio element is used to output the embedded speech or synthesized speech of an audio file. A src attribute can be designated in the audio element. The Uniform Resource Identifier (URI) of the audio file is designated in the src attribute. For example, in Fig. 19, the text information "What city do you want to fly from?" is read aloud by the reproduction of the audio file "prompt.au" designated in the src attribute.

[0126]

Using the audio element allows for the reproduction, for example, of a recorded audio file, and thus allows for provision of the speech information as the producer of the

user interface intends to people with visual disabilities.
[0127]

Note that the sub element, phoneme element, and audio
element described above are examples of description of
5 text-to-speech utterance metadata in the SSML format.
Another element or attribute in the SSML format can be used
for the description. Alternatively, the text-to-speech
utterance metadata can be described, for example, in another
markup language other than the SSML format.

10 [0128]

<5. Flow of Process Performed in Each Device>

[0129]

Next, the processes performed in the transmission device
10 and reception device 20 included in the broadcasting system
15 1 of Fig. 7 will be described.

[0130]

(Transmitting Process)

First, the flow of a transmitting process that the
transmission device 10 of Fig. 7 performs will be described
20 with reference to the flowchart of Fig. 20.

[0131]

In step S111, the content acquisition unit 111 acquires
the content, for example, of a broadcast program and provides
the content to the stream generation unit 114.

25 [0132]

In step S112, the text-to-speech utterance metadata
generation unit 112 generates text-to-speech utterance
metadata, for example, in response to the instructions from
the producer of the user interface, and provides the
30 text-to-speech utterance metadata to the ESG information
generation unit 113.

[0133]

Note that when the text-to-speech utterance metadata generation unit 112 generates text-to-speech utterance metadata including address information and provides the text-to-speech utterance metadata to the ESG information generation unit 113, the text-to-speech utterance metadata generation unit 112 generates a text-to-speech utterance metadata file to be acquired in accordance with the address information, and provides the text-to-speech utterance metadata file to the stream generation unit 114.

[0134]

In step S113, the ESG information generation unit 113 generates ESG information on the basis of the text-to-speech utterance metadata provided from the text-to-speech utterance metadata generation unit 112, and provides the ESG information to the stream generation unit 114.

[0135]

In step S114, the stream generation unit 114 generates a stream in compliance with a predetermined standard on the basis of the content data provided from the content acquisition unit 111 and the ESG information provided from the ESG information generation unit 113, and provides the stream to the transmission unit 115.

[0136]

Note that, when the text-to-speech utterance metadata included in the ESG information provided from the ESG information generation unit 113 includes the address information, the stream generation unit 114 generates a stream in compliance with a predetermined standard on the basis of the text-to-speech utterance metadata file provided from the text-to-speech utterance metadata generation unit 112 in

addition to the content data and the ESG information, and provides the stream to the transmission unit 115.

[0137]

In step S115, the transmission unit 115 processes the stream provided from the stream generation unit 114, for example, in a digital modulation process, and transmits the stream as a digital broadcast signal through the antenna 116.

[0138]

The transmitting process has been described above. In the transmitting process, the text-to-speech utterance metadata about the text-to-speech utterance of the display information, for example, of a user interface as the producer intends is generated, and the ESG information including the text-to-speech utterance metadata is generated. Then, the ESG information including the text-to-speech utterance metadata is transmitted together with the content.

[0139]

This process enables the TTS engine 219 on the reception device 20 to read the display information aloud on the basis of the text-to-speech utterance metadata. Thus, for example, even when the pronunciation of the text information is not uniquely determined, or the text information is a proper noun difficult to pronounce, the text information is surely read aloud as the producer intends. As a result, people with visual disabilities can receive the information similar to the information that able-bodied people receive.

[0140]

(Receiving Process)

Next, the flow of a receiving process that the reception device 20 of Fig. 7 performs will be described with reference to the flowchart of Fig. 21.

[0141]

In step S211, the reception unit 212 receives the digital broadcast signal transmitted through an antenna 211 from the transmission device 10. Furthermore, the reception unit 212
5 processes the digital broadcast signal, for example, in a demodulation process, and provides the stream generated by the process to the stream separation unit 213.

[0142]

In step S212, the stream separation unit 213 separates
10 the content data and the ESG information from the stream provided from the reception unit 212, and provides the content data to the reproduction unit 214, and the ESG information to the ESG information acquisition unit 217.

[0143]

15 In step S213, the ESG information acquisition unit 217 acquires the ESG information provided from the stream separation unit 213. For example, when the user instructs the reception device 20 to display an ESG screen, the ESG information acquisition unit 217 provides the ESG information
20 to the reproduction unit 214. Furthermore, the ESG information acquisition unit 217 provides the text-to-speech utterance metadata included in the ESG information to the text-to-speech utterance metadata acquisition unit 218.

[0144]

25 In step S214, the reproduction unit 214 generates an ESG screen on the basis of the ESG information provided from the ESG information acquisition unit 217, and displays the ESG screen on the display unit 215.

[0145]

30 In step S215, the text-to-speech utterance metadata acquisition unit 218 acquires the text-to-speech utterance

metadata provided from the ESG information acquisition unit 217.

[0146]

When the text-to-speech utterance metadata includes the address information in this step, the text-to-speech utterance metadata acquisition unit 218 acquires the text-to-speech utterance metadata file from the stream separated by the stream separation unit 213 on the basis of the address information, and provides the text-to-speech utterance metadata including the content provided from the acquired text-to-speech utterance metadata file to the TTS engine 219. On the other hand, when the text-to-speech utterance metadata includes the content, the text-to-speech utterance metadata acquisition unit 218 provides the text-to-speech utterance metadata without any process to the TTS engine 219.

[0147]

In step S216, the TTS engine 219 reads the display information, for example, of a user interface aloud on the basis of the text-to-speech utterance metadata provided from the text-to-speech utterance metadata acquisition unit 218, and outputs the sound of the display information from the loudspeaker 216.

[0148]

In this example, when an ESG screen is displayed on the display unit 215 and a title or program information is to be read aloud in order to make the ESG screen accessible for people with visual disabilities, but, for example, the pronunciation of the text information is not uniquely determined in the process of step S214, the TTS engine 219 follows the text-to-speech utterance metadata to read the text information aloud as the producer intends.

[0149]

The receiving process has been described above. In the receiving process, the ESG information, which includes the text-to-speech utterance metadata about the text-to-speech utterance of the display information as the producer intends and is transmitted from the transmission device 10, is received. Subsequently, the text-to-speech utterance metadata included in the ESG information is acquired. Then, the display information, for example, of a user interface is read aloud on the basis of the text-to-speech utterance metadata.

[0150]

This process enables the TTS engine 219 to read the display information aloud on the basis of the text-to-speech utterance metadata. Thus, for example, when the pronunciation of the text information is not uniquely determined, or the text information is a proper noun difficult to pronounce, the text information is surely read aloud as the producer intends. As a result, people with visual disabilities can receive the information similar to the information that able-bodied people receive.

[0151]

<6. Exemplary Variation>

[0152]

When the text-to-speech utterance metadata includes the address information, the text-to-speech utterance metadata file transmitted in the FLUTE session is acquired in accordance with the address information in the embodiment described above. However, the text-to-speech utterance metadata file can be delivered from a server on the Internet. In such a case, for example, the Uniform Resource Locator (URL) of the server is designated as the address information.

[0153]

The ESG prescribed in OMA-BCAST has been described as the electronic program information in the embodiment described above. The present technology can be applied, for example, to Electronic Program Guide (EPG) or another electronic program information. Furthermore, the electronic program information such as ESG information can be delivered from a server on the Internet and received by the reception device 20.

10 [0154]

<7. Configuration of Computer>

[0155]

The series of processes described above can be performed either with hardware or with software. When the series of processes is performed with software, the program of the software is installed on the computer. Fig. 22 is a diagram of exemplary configuration of the hardware of the computer that performs the series of processes with a program.

[0156]

20 In a computer 900, a Central Processing Unit (CPU) 901, a Read Only Memory (ROM) 902, and a Random Access Memory (RAM) 903 are connected to each other via a bus 904. An input/output interface 905 is further connected to the bus 904. An input unit 906, an output unit 907, a recording unit 908, a communication unit 909, and a drive 910 are connected to the input/output interface 905.

[0157]

30 The input unit 906 includes, for example, a keyboard, a mouse, and a microphone. The output unit 907 includes, for example, a display, and a loudspeaker. The recording unit 908 includes, for example, a hard disk or a non-volatile memory.

The communication unit 909 includes, for example, a network interface. The drive 910 drives a removable medium 911 such as a magnetic disk, an optical disk, a magneto-optical disk, or a semiconductor memory.

5 [0158]

In the computer 900 having the configuration described above, the CPU 901 loads the program stored in the ROM 902 or the recording unit 908 via the input/output interface 905 and the bus 904 onto the RAM 903 and executes the program.

10 This execution implements the series of processes.

[0159]

The program executed by the computer 900 (CPU 901) can be recorded in the removable medium 911, for example, as a package medium and provided. Alternatively, the program can
15 be provided through a wired or wireless transmission medium such as a local area network, the Internet, or digital satellite broadcasting.

[0160]

The computer 900 can install a program via the
20 input/output interface 905 onto the recording unit 908 by attaching the removable medium 911 to the drive 910.

Alternatively, the program can be received in the communication unit 909 via a wired or wireless transmission medium, and installed on the recording unit 908. Alternatively, the
25 program can previously be installed on the ROM 902 or the recording unit 908.

[0161]

Herein, the processes that the computer performs in accordance with a program are not necessarily performed
30 chronologically in the order described in the flowchart. In other words, the processes that the computer performs in

accordance with a program include processes performed in parallel, or individually (for example, parallel processing or object processing). Furthermore, the program can be executed by a computer (processor), or by the decentralized processing with a plurality of computers.

[0162]

Note that the embodiments of the present technology are not limited to the embodiments describe above, and can variously be changed without departing from the gist of the present technology.

[0163]

Alternatively, the present technology can have the following configuration.

[0164]

(1)

A transmission device including:

a metadata generation unit that generates metadata about text-to-speech utterance of display information, the text-to-speech utterance being performed as a producer intends;

an electronic program information generation unit that generates electronic program information including the metadata; and

a transmission unit that transmits the electronic program information to a reception device capable of displaying the display information.

(2)

The transmission device according to (1), wherein the metadata includes information about utterance of a character string of which pronunciation is not uniquely determined, or a character string difficult to pronounce.

(3)

The transmission device according to (1) or (2), wherein the display information includes information or an icon of a content.

5 (4)

The transmission device according to (3), further including:

a content acquisition unit that acquires the content, wherein the transmission unit transmits the electronic
10 program information together with the content using a digital broadcast signal.

(5)

The transmission device according to any of (1) to (4), wherein

15 the electronic program information complies with Electronic Service Guide (ESG) prescribed in Open Mobile Alliance - Mobile Broadcast Services Enabler Suite (OMA-BCAST),

the metadata is described in a Speech Synthesis Markup
20 Language (SSML) format, and

a predetermined fragment included in the ESG includes address information indicating an address at which a file of the metadata described in the SSML format is acquired, or content of the metadata described in the SSML format.

25 (6)

A transmission method for causing a transmission device to perform a process, the process including:

generating metadata about text-to-speech utterance of display information, the text-to-speech utterance being
30 performed as a producer intends;

generating electronic program information including

the metadata; and

transmitting the electronic program information to a reception device capable of displaying the display information.

5 (7)

A reception device including:

a reception unit that receives electronic program information, the electronic program information being transmitted from a transmission device, the electronic program
10 information including the metadata about text-to-speech utterance of display information, the text-to-speech utterance being performed as a producer intends;

a metadata acquisition unit that acquires the metadata included in the electronic program information; and

15 a text-to-speech reading aloud unit that reads the display information aloud on the basis of the metadata.

(8)

The reception device according to (7), wherein the metadata includes information about utterance of a character
20 string of which pronunciation is not uniquely determined, or a character string difficult to pronounce.

(9)

The reception device according to (7) or (8), wherein the display information includes information or an icon of
25 a content.

(10)

The reception device according to (9), wherein the reception unit receives the electronic program information transmitted together with the content as a digital broadcast
30 signal.

(11)

The reception device according to any of (7) to (10),
wherein

the electronic program information complies with ESG
prescribed in OMA-BCAST,

5 the metadata is described in an SSML format,
a predetermined fragment included in the ESG includes
address information indicating an address at which a file of
the metadata described in the SSML format is acquired, or
content of the metadata described in the SSML format, and
10 the metadata acquisition unit acquires the file of the
metadata in accordance with the address information, or
acquires the metadata from the fragment.

(12)

A reception method for causing a reception device to
15 perform a process, the process including:

receiving electronic program information, the
electronic program information being transmitted from a
transmission device, the electronic program information
including the metadata about text-to-speech utterance of
20 display information, the text-to-speech utterance being
performed as a producer intends;

acquiring the metadata included in the electronic
program information; and

25 reading the display information aloud on the basis of
the metadata.

REFERENCE SIGNS LIST

[0165]

1 Broadcasting system
30 10 Transmission device
20 Reception device

	111	Content acquisition unit
	112	Text-to-speech utterance metadata generation unit
	113	ESG information generation unit
	114	Stream generation unit
5	115	Transmission unit
	212	Reception unit
	213	Stream separation unit
	214	Reproduction unit
	215	Display unit
10	216	Loudspeaker
	217	ESG information acquisition unit
	218	Text-to-speech utterance metadata acquisition unit
	219	TTS engine
	900	Computer
15	901	CPU

CLAIMS

1. A transmission device comprising:
 - ametadata generation unit that generates metadata about
 - 5 text-to-speech utterance of display information, the
 - text-to-speech utterance being performed as a producer
 - intends;
 - an electronic program information generation unit that
 - generates electronic program information including the
 - 10 metadata; and
 - a transmission unit that transmits the electronic
 - program information to a reception device capable of displaying
 - the display information.
- 15 2. The transmission device according to claim 1, wherein
- the metadata includes information about utterance of a
- character string of which pronunciation is not uniquely
- determined, or a character string difficult to pronounce.
- 20 3. The transmission device according to claim 2, wherein
- the display information includes information or an icon of
- a content.
4. The transmission device according to claim 3, further
- 25 comprising:
 - a content acquisition unit that acquires the content,
 - wherein the transmission unit transmits the electronic
 - program information together with the content using a digital
 - broadcast signal.
- 30 5. The transmission device according to claim 1, wherein

the electronic program information complies with
Electronic Service Guide (ESG) prescribed in Open Mobile
Alliance - Mobile Broadcast Services Enabler Suite
(OMA-BCAST),

5 the metadata is described in a Speech Synthesis Markup
Language (SSML) format, and

 a predetermined fragment included in the ESG includes
address information indicating an address at which a file of
the metadata described in the SSML format is acquired, or
10 content of the metadata described in the SSML format.

6. A transmission method for causing a transmission device
to perform a process, the process comprising:

 generating metadata about text-to-speech utterance of
15 display information, the text-to-speech utterance being
performed as a producer intends;

 generating electronic program information including
the metadata; and

 transmitting the electronic program information to a
20 reception device capable of displaying the display
information.

7. A reception device comprising:

 a reception unit that receives electronic program
25 information, the electronic program information being
transmitted from a transmission device, the electronic program
information including the metadata about text-to-speech
utterance of display information, the text-to-speech
utterance being performed as a producer intends;

30 a metadata acquisition unit that acquires the metadata
included in the electronic program information; and

a text-to-speech reading aloud unit that reads the display information aloud on the basis of the metadata.

8. The reception device according to claim 7, wherein the
5 metadata includes information about utterance of a character string of which pronunciation is not uniquely determined, or a character string difficult to pronounce.

9. The reception device according to claim 8, wherein the
10 display information includes information or an icon of a content.

10. The reception device according to claim 9, wherein the reception unit receives the electronic program information
15 transmitted together with the content as a digital broadcast signal.

11. The reception device according to claim 7, wherein
the electronic program information complies with ESG
20 prescribed in OMA-BCAST,

the metadata is described in an SSML format,
a predetermined fragment included in the ESG includes address information indicating an address at which a file of the metadata described in the SSML format is acquired, or
25 content of the metadata described in the SSML format, and
the metadata acquisition unit acquires the file of the metadata in accordance with the address information, or acquires the metadata from the fragment.

30 12. A reception method for causing a reception device to perform a process, the process comprising:

receiving electronic program information, the
electronic program information being transmitted from a
transmission device, the electronic program information
including the metadata about text-to-speech utterance of
5 display information, the text-to-speech utterance being
performed as a producer intends;

acquiring the metadata included in the electronic
program information; and

reading the display information aloud on the basis of
10 the metadata.