



(11)

EP 4 826 710 A1

(12)

EUROPEAN PATENT APPLICATION
published in accordance with Art. 153(4) EPC

(43) Date of publication:
21.10.2026 Bulletin 2026/43

(51) International Patent Classification (IPC):
G10G 3/04 (2006.01)

(21) Application number: **24903212.9**

(52) Cooperative Patent Classification (CPC):
G10G 3/04

(22) Date of filing: **23.07.2024**

(86) International application number:
PCT/JP2024/026226

(87) International publication number:
WO 2025/126548 (19.06.2025 Gazette 2025/25)

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC ME MK MT NL NO PL PT RO RS SE SI SK SM TR
Designated Extension States:
BA
Designated Validation States:
GE KH MA MD TN

(71) Applicant: **Sony Group Corporation**
Tokyo 108-0075 (JP)

(72) Inventors:
• **SATO, Gakusei**
Tokyo 141-0022 (JP)
• **AKAMA, Taketo**
Tokyo 141-0022 (JP)

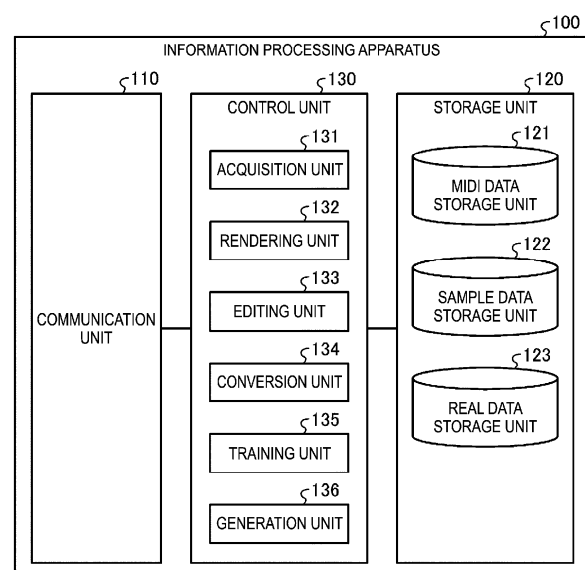
(30) Priority: **13.12.2023 US 202363609496 P**

(74) Representative: **2SPL Patentanwälte PartG mbB**
Landaubogen 3a
81373 München (DE)

(54) **INFORMATION PROCESSING SYSTEM, INFORMATION PROCESSING METHOD, AND INFORMATION PROCESSING PROGRAM**

(57) Proposed is an information processing system, an information processing method, and an information processing program that enable automatic generation of general-purpose and highly accurate transcription data. An information processing system according to the present disclosure includes an acquisition unit configured to acquire audio data to be processed; and a generation unit configured to generate corresponding transcription data by inputting the audio data acquired by the acquisition unit into a trained machine learning model trained using sample data of audio.

Fig. 6



Description

Technical Field

[0001] The present disclosure relates to an information processing system, an information processing method, and an information processing program.

Background Art

[0002] In recent years, technology for automatically generating transcription data from audio such as music has been attracting attention. Such technology may be called Automatic Music Transcription (AMT) and is an important technique in music information processing.

[0003] As technology related to AMT, technology is known in which a machine learning model is trained using a set of annotation data and audio data that correspond to audio as teacher data to enable automatic generation of highly accurate transcription data for musical instruments for which such teacher data is abundant (for example, NPL 1 to NPL 4).

Citation List

Non Patent Literature

[0004]

NPL 1: J. Gardner et al. "MT3: multi-task multitrack music transcription" in Proceedings of ICLR, 2021

NPL 2: S. Ian et al. "Scaling polyphonic transcription with mixtures of monophonic transcriptions" in Proceedings of ISMIR, 2022

NPL 3: R. M. Bittner et al. "A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation" in Proceedings of ICASSP, 2022

NPL 4: Y.-T. Wu et al. "Multi-instrument automatic music transcription with self-attention-based instance segmentation" in Proceedings of ICML, 2022

Summary

Technical Problem

[0005] However, with the known technology, when such teacher data is poor or unavailable, automatic generation of highly accurate transcription data cannot be performed.

[0006] Thus, the present disclosure proposes an information processing system, an information processing method, and an information processing program that enable automatic generation of general-purpose and highly accurate transcription data.

Solution to Problem

[0007] To solve the problems described above, an information processing system according to an aspect of the present disclosure includes an acquisition unit configured to acquire audio data to be processed; and a generation unit configured to generate corresponding transcription data by inputting the audio data acquired by the acquisition unit into a trained machine learning model trained using sample data of audio.

Brief Description of Drawings

[0008]

[Fig. 1]

Fig. 1 is a diagram illustrating an overview of an information processing system according to an embodiment.

[Fig. 2]

Fig. 2 is a diagram illustrating an example of MIDI data according to an embodiment.

[Fig. 3]

Fig. 3 is a diagram illustrating an overview of application processing using a machine learning model according to an embodiment.

[Fig. 4]

Fig. 4 is a diagram illustrating an example of experimental results acquired using a proposed method according to an embodiment.

[Fig. 5]

Fig. 5 is a diagram illustrating a configuration example of a user terminal according to an embodiment.

[Fig. 6]

Fig. 6 is a diagram illustrating a configuration example of an information processing apparatus according to an embodiment.

[Fig. 7]

Fig. 7 is a diagram illustrating an example of a MIDI data storage unit according to an embodiment.

[Fig. 8]

Fig. 8 is a diagram illustrating an example of a sample data storage unit according to an embodiment.

[Fig. 9]

Fig. 9 is a diagram illustrating an example of a real data storage unit according to an embodiment.

[Fig. 10]

Fig. 10 is a flowchart illustrating a flow of learning processing in a control unit.

[Fig. 11]

Fig. 11 is a flowchart illustrating a flow of application processing in a control unit.

[Fig. 12]

Fig. 12 is a diagram of a hardware configuration illustrating an example of a computer that realizes a function of the information processing apparatus.

Description of Embodiments

[0009] Hereinafter, embodiments of the present disclosure will be described in detail with reference to the drawings. Note that, in the following embodiments, the same reference signs are assigned to identical components, and redundant descriptions will be omitted accordingly.

[0010] The present disclosure will be described in accordance with the following itemized order.

1. Embodiments

1-1. Overview of Information Processing Apparatus according to Embodiment

1-2. Overview of Information Processing System according to Embodiment

1-3. Configuration of Information Processing Apparatus according to Embodiment

2. Other Embodiments

3. Effects of Information Processing Apparatus according to Present Disclosure

4. Hardware Configuration

1. Embodiments

1-1. Overview of Information Processing Apparatus according to Embodiment

[0011] First, an overview of an information processing apparatus 100 according to an embodiment will be described. The information processing apparatus 100 is an information processing apparatus designed to automatically generate general-purpose and highly accurate transcription data and may be any apparatus as long as the apparatus can realize the processing according to the embodiment. The information processing apparatus 100 is implemented by, for example, a server apparatus, a cloud system, or the like, and executes information processing according to the embodiment.

[0012] For example, the information processing apparatus 100 uses a machine learning model that has been trained using audio sample data. Further, for example, the information processing apparatus 100 executes various types of data processing on audio data transmitted from an external information processing apparatus (such as a user terminal 10 described later) and provides transcription data of audio desired by the user.

[0013] The information processing apparatus 100 uses audio sample data as audio data for training the machine learning model. For example, the information processing apparatus 100 uses sample data of a plurality of pieces of audio called "One Shot" in the Musical Instruments Digital Interface (MIDI) standard.

[0014] "One Shot" is predetermined audio data and is

audio data of a sound divided into certain units, such as a sound of "do" and a sound of "re". For example, the audio data is audio data of each sound such as a sound of "do" or a sound of "re". Then, the information processing apparatus 100 uses the synthetic data acquired by synthesizing such sample data as audio data for when training the machine learning model.

[0015] The information processing apparatus 100 performs training using two types of methods for the learning method of the machine learning model. In the first method, the information processing apparatus 100 uses MIDI data. The MIDI data includes note information relating to the onset time, offset time, pitch speed, and the like of notes, and the information processing apparatus 100 acquires these pieces of note information and synthesizes them by rendering the pieces of note information with the sample data.

[0016] That is, the information processing apparatus 100 performs additive synthesis of sample data according to the MIDI data. Thus, audio data corresponding to the MIDI data is generated. That is, a set of audio data and MIDI data is generated. The set of the audio data and the MIDI data is the teacher data according to the embodiment. The information processing apparatus 100 then trains the machine learning model using the synthetic data that is the synthesizing result.

[0017] In the second method, the information processing apparatus 100 acquires real data and trains the machine learning model so as to reduce the divergence from the real data. By training the machine learning model so as to reduce the divergence, even when the audio data to be processed is real audio data, it is possible to automatically generate highly accurate transcription data.

[0018] By using these two types of learning methods, even when such teacher data is scarce or unavailable, general-purpose and highly accurate automatic generation of transcription data can be performed. Further, even when the learning methods are applied to a domain (such as a musical instrument (including a way of playing), a tone, a singer (including a way of singing), a genre, or a language) in which such teacher data is scarce or unavailable, it is possible to automatically generate general-purpose and highly accurate transcription data.

1-2. Overview of Information Processing System according to Embodiment

[0019] Next, an overview of an information processing system for describing the technology of the present disclosure will be described with reference to Fig. 1. In the following embodiment, an example is described in which the audio data at the time of application of the machine learning model is from a piano and the audio transcription data desired by a user U1 is piano keystroke information, but the audio data is not limited to that of a musical instrument such as a piano.

For example, the audio data may be singing data or the like.

[0020] Fig. 1 is a diagram illustrating an overview of an information processing system according to the embodiment. In Fig. 1, the machine learning model according to the embodiment is described as a machine learning model M1. The machine learning model M1 includes an encoder 20 and a decoder 30. The encoder 20 performs learning in cooperation with a discriminator 40. The encoder 20 receives a feature amount Xs acquired by encoding synthetic data D1 and a feature amount Xr acquired by encoding real data D2.

[0021] The discriminator 40 also determines whether the output information from the encoder 20 is the synthetic data D1 or the real data D2 (determination H1). For example, the discriminator 40 includes three fully connected layers. Further, for example, the discriminator 40 receives information acquired by performing section selection on the output information from the encoder 20 and flattening the output information.

[0022] Here, the learning processing of the machine learning model M1 will be described. The information processing apparatus 100 acquires MIDI data F1 and sample data F2. The three bars included in the MIDI data F1 are information related to the tone and are information defining details for audio reproduction, such as which note the tone is based on, when the note starts and ends, and at what pitch speed the note is played. This information included in the MIDI data F1 is only an example, and in actuality, the information is complex information for playing.

[0023] Here, the MIDI data F1 will be described with reference to Fig. 2. Fig. 2 is a diagram illustrating an example of MIDI data according to the embodiment. The MIDI data illustrated in Fig. 2 is data with the pitch (corresponding to keys in the case of a piano) and time as axes. The alphanumeric characters such as "F3", "B3", "D4", and "F4" on the pitch axis are information indicating the pitch.

[0024] In Fig. 2, note N1 to note N4 are arranged side by side at the beginning of the data, which indicates that a plurality of pitches are played simultaneously. Specifically, this indicates that the pitches "G3#", "B3", "D4", and "D5" corresponding to the notes N1 to N4 are simultaneously played. For example, in the case of a piano, it indicates that the keys corresponding to "G3#", "B3", "D4", and "D5" are simultaneously played. Note that, in Fig. 2, only the first four bars of the data are denoted by reference numerals, but this is for convenience of description, and reference numerals may be appropriately assigned to other bars included in the data.

[0025] The description returns to Fig. 1. The sample data F2 includes a plurality of audio sample data. For example, sample data F21 to sample data F23 are included. The sample data F2 includes a plurality of types of audio data such as main audio Z1 and sub audio Z2. The sample data F2 includes the main audio

Z1 and the sub audio Z2 corresponding to each of the sample data F21 to sample data F23.

[0026] The information included in the sample data F2 is an example, and a larger number of pieces of sample data are actually included. For example, although not illustrated, sample data F24, sample data F25, ..., and the like may be included. The sample data F2 may include only the main audio Z1 (without the sub audio Z2). The sample data F2 may include audio data of a type different from the main audio Z1 and the sub audio Z2. For example, the sub audio Z2 may further include audio data of the sub audio type.

[0027] The sample data F2 may include sample data of a main audio type other than the main audio Z1 or may include sample data of a sub audio type other than the sub audio Z2. The main audio Z1 may be made of a set of sample data of a plurality of main audio types, and the sub audio Z2 may be made of a set of sample data of a plurality of sub audio types.

[0028] The information processing apparatus 100 may generate new "One Shot" audio data by combining a plurality of types of audio data. For example, the information processing apparatus 100 may generate new "One Shot" audio data by combining audio data of the plurality of types of the main audio Z1 and the sub audio Z2.

[0029] For example, the information processing apparatus 100 may weight the main audio Z1 and the sub audio Z2 and generate new "One Shot" audio data in accordance with the weighting. For example, since the sound of "do" of the piano and the sound of "do" of the guitar are different even though the sound of "do" is the same, the information processing apparatus 100 may generate "One Shot" audio data of a new sound of "do" by weighting each of the sounds. Accordingly, the information processing apparatus 100 can generate new "One Shot" audio data including both elements of the piano and the guitar. As a result, it is possible to generate the synthetic data D1 of various tones.

[0030] In addition, in order to achieve a variety, the information processing apparatus 100 may generate the synthetic data D1 while changing the weighting so that the tone is slightly changed in accordance with the music in the time-series direction. For example, the information processing apparatus 100 may generate the synthetic data D1 by using completely different musical instruments while changing the weighting for each musical instrument so that the tone changes in the time-series direction.

[0031] In addition, the information processing apparatus 100 is not limited to the case of synthesizing two types of audio data, and in a case where there are a plurality of types of main audio and a plurality of types of sub audio, the information processing apparatus 100 may generate new "One Shot" audio data by appropriately combining the plurality of types of audio data. For example, the information processing apparatus 100 may generate new "One Shot" audio data by combining

the main audio Z1, the sub audio Z2, and the sub audio Z3.

[0032] Further, the information processing apparatus 100 may generate new "One Shot" audio data by combining sample data of the same musical instrument or the like with different types of audio data or may generate new "One Shot" audio data by combining sample data of different musical instruments or the like with different types of audio data. For example, the information processing apparatus 100 may combine the sound of a piano and the sound of a piano (for example, a case of combining the sound of a high-grade grand piano of the brand "OO" and the sound of a non-high-grade jazz piano of the brand "O×"), may combine the sound of a piano and the sound of a guitar, or may combine the sound of a piano and a singing voice.

[0033] The information processing apparatus 100 renders the sample data F2 and the MIDI data F1. Specifically, the information included in the sample data F2 and the information included in the MIDI data F1 are rendered to generate the synthetic data D1. For example, when the information processing apparatus 100 acquires information indicating that the tone of the first bar is the sound of "do" of a piano and the time is one second from "0:01 to 0:02" (the time from the key being played to the release being one second) from the MIDI data F1, the information processing apparatus 100 identifies the audio corresponding to the tone of the bar from the sample data F2 and renders the audio.

[0034] By rendering the sample data F2 and the MIDI data F1 in this manner, the audio of the sample data F2 can be reproduced in accordance with the MIDI data F1. As a result, it is possible to automatically generate a set of the audio data (synthetic data D1) acquired by rendering the sample data F2 and the MIDI data F1 and the MIDI data F1.

[0035] Note that the information processing apparatus 100 may execute predetermined editing processing R1 (preprocessing for training the machine learning model M1) when generating the synthetic data D1. For example, the information processing apparatus 100 may execute the editing processing R1 including compression or limiting. This is because, if the rendered audio is used as is, the audio may just be audio acquired by linearly adding the sounds of the notes of the MIDI data F1.

[0036] Thus, the editing processing R1 including compressing dynamics may be executed so that the volume is constant to some extent when one note is reproduced in the track and when a plurality of notes are reproduced at the same time. In addition, the editing processing R1 including applying limiting based on a predetermined value may be executed so that the volume is constant to some extent (cutting off a portion exceeding the predetermined threshold as an upper limit of the amplitude).

[0037] The information processing apparatus 100 acquires the synthetic data D1 and the real data D2. In

this case, the information processing apparatus 100 may acquire, for example, the real data D2 corresponding to the synthetic data D1 (for example, the predetermined real data D2 associated with the MIDI data F1) or may acquire the real data D2 in accordance with the actual AMT use (for example, the real data D2 close to the audio data to be processed).

[0038] In the latter case, for example, the machine learning model M1 may be tuned for each of the users via training using the real data D2 corresponding to the actual use of the AMT for each of the users. Thus, for example, the information processing apparatus 100 may acquire, from the user U1, the real data D2 corresponding to (e.g., similar to) the real audio data that the user U1 wants to transcribe. For example, when the data acquired from the user U1 is insufficient for training as the real data D2 corresponding to the real audio data that the user U1 desires to transcribe, the information processing apparatus 100 may search for the real data D2 corresponding to this data and acquire the real data D2 from an external information processing apparatus or the like.

[0039] The information processing apparatus 100 then acquires the feature amount Xs acquired by encoding the synthetic data D1 and the feature amount Xr acquired by encoding the real data D2. Specifically, the information processing apparatus 100 converts the synthetic data D1 into the feature amount Xs by encoding the synthetic data D1 and converts the real data D2 into the feature amount Xr by encoding the real data D2, thus acquiring the feature amount Xs and the feature amount Xr.

[0040] With known technology, when the feature amount Xs is input to the encoder 20, the encoder 20 is trained to output the MIDI data F1, which is the original data of the feature amount Xs. However, when the audio data to be processed is real audio data, the accuracy of automatic generation of the transcription data may be low.

[0041] The information processing apparatus 100 inputs the acquired feature amounts (the feature amount Xs and the feature amount Xr) to the encoder 20. The output information from the encoder 20 is input to the discriminator 40. The discriminator 40 performs a determination (the determination H1) as to whether the output information corresponding to the synthetic data D1 and the output information corresponding to the real data D2 are the synthetic data D1 or the real data D2. Then, training is performed based on the output information from the discriminator 40.

[0042] Note that a loss function such as a Binary Cross Entropy Loss (BCE Loss) is used for learning. The information processing apparatus 100 performs two types of learning using such a loss function. The two types of learning are a "Classification Loss" L1 and an "Adversarial Loss" L2. First, with the "Classification Loss" L1, the information processing apparatus 100 calculates a simple loss with respect to the correct data

for the discriminator 40 and updates the parameters using the calculation result so as to improve the performance of the discriminator 40. In this case, the information processing apparatus 100 updates the parameter based on the loss between the information acquired from the real data D2 and the information acquired from the synthetic data D1, using the information acquired from the real data D2 as the correct data.

[0043] In this manner, in the "Classification Loss" L1, the encoder 20 is fixed and training is performed for the discriminator 40. In Fig. 1, $BCE(0, C(E(X_s)))$ is information corresponding to the synthetic data D1, and $BCE(1, C(E(X_r)))$ is information corresponding to the real data D2. Here, the numbers "0" and "1" in parentheses of the BCE indicate that learning is performed so as to approach "0" or "1". For example, $BCE(0, C(E(X_s)))$ indicates that learning is performed so that information acquired from $C(E(X_s))$ approaches "0", and $BCE(1, C(E(X_r)))$ indicates that learning is performed so that information acquired from $C(E(X_r))$ approaches "1".

[0044] The information processing apparatus 100 performs learning so that information acquired from the synthetic data D1 via the encoder 20 and the discriminator 40 approaches "0" and performs learning so that information acquired from the real data D2 via the encoder 20 and the discriminator 40 approaches "1".

[0045] On the other hand, with "Adversarial Loss", the information processing apparatus 100 performs adversarial learning. Specifically, the information processing apparatus 100 performs learning so that the synthetic data D1 and the real data D2 cannot be distinguished from one another. More specifically, the information processing apparatus 100 performs learning so as to reduce divergence between the information corresponding to the synthetic data D1 and the information corresponding to the real data D2.

[0046] For such adversarial learning, the information processing apparatus 100 backpropagates the loss for the encoder 20 to update the parameters. That is, the information processing apparatus 100 updates the parameters by performing the adversarial learning by backpropagating the loss for the encoder 20 so that the discriminator 40 has difficulty in discriminating between the synthetic data D1 and the real data D2. As described above, in the "Adversarial Loss" L2, contrary to the "Classification Loss" L1, the discriminator 40 is fixed and training is performed for the encoder 20.

[0047] In Fig. 1, $BCE(0.5, C(E(X_s)))$ is information corresponding to the synthetic data D1, and $BCE(0.5, C(E(X_r)))$ is information corresponding to the real data D2. The information processing apparatus 100 changes the learning condition from $BCE(0, C(E(X_s)))$ to $BCE(0.5, C(E(X_s)))$ and from $BCE(1, C(E(X_r)))$ to $BCE(0.5, C(E(X_r)))$ (changes the learning condition to 0.5, which is an intermediate value between 0 and 1) to perform learning so that the synthetic data D1 and the real data D2 cannot be distin-

guished from one another.

[0048] The information processing apparatus 100 performs learning so that information acquired from the synthetic data D1 via the encoder 20 and the discriminator 40 approaches "0.5" and performs learning so that information acquired from the real data D2 via the encoder 20 and the discriminator 40 approaches "0.5".

[0049] In addition to or instead of learning that makes it impossible to distinguish between the synthetic data D1 and the real data D2, the information processing apparatus 100 may perform learning so that the synthetic data D1 is mistaken as the real data D2. The information processing apparatus 100 may change the learning condition from $BCE(0, C(E(X_s)))$ to $BCE(1, C(E(X_s)))$ and from $BCE(1, C(E(X_r)))$ to $BCE(0, C(E(X_r)))$, or may change the learning condition from $BCE(0.5, C(E(X_s)))$ to $BCE(1, C(E(X_s)))$ and from $BCE(0.5, C(E(X_s)))$ to $BCE(0, C(E(X_r)))$ to perform learning so that the synthetic data D1 is mistaken as the real data D2.

[0050] The information processing apparatus 100 performs learning so that information acquired from the synthetic data D1 via the encoder 20 and the discriminator 40 approaches "1" and performs learning so that information acquired from the real data D2 via the encoder 20 and the discriminator 40 approaches "0". Hereinafter, a learning method that makes it impossible to distinguish between the synthetic data D1 and the real data D2 will be referred to as "Domain Confusion", and a learning method that makes the synthetic data D1 mistaken for the real data D2 will be referred to as "Domain Adaptation".

[0051] Through the two types of learning ("Classification Loss" L1 and "Adversarial Loss" L2), automatic generation of highly accurate transcription data can be achieved. As described above, the discriminator 40 is incorporated after the encoder 20, and learning is performed so that the synthetic data D1 and the real data D2, which are identified as different pieces of data at the time of input, are not identified as different pieces of data at the time of output. Thus, highly accurate and appropriate processing can be executed even when the audio data to be processed is real audio data.

[0052] Then, the information processing apparatus 100 generates the transcription data via the encoder 20 and the decoder 30 as with known technology. The information processing apparatus 100 generates output information F3 (note information relating to pitch, time, pitch speed, and the like) such as the MIDI data F1 output from the decoder 30 via the encoder 20 and the decoder 30 to generate transcription data. In this case, as with known technology, the information processing apparatus 100 learns "Transcription Loss" L3 using the input information of the encoder 20 and the output information F3 of the decoder 30 and performs optimization for generating the transcription data.

[0053] The output information F3 output from the decoder 30 may be transcription data. That is, tran-

scription data may be directly output from the machine learning model M1 or information (information such as the MIDI data F1) for generating transcription data may be output. In the latter case, the information processing apparatus 100 generates transcription data based on information (information such as the MIDI data F1) output from the machine learning model M1.

[0054] The learning processing of the machine learning model M1 has been described above. Here, the application processing using the trained machine learning model M1 will be described with reference to Fig. 3. Fig. 3 is a diagram illustrating an example of application processing using a machine learning model according to the embodiment. Before describing Fig. 3, the user terminal 10 will be described first.

[0055] The user terminal 10 is an information processing apparatus used by a user who desires transcription data corresponding to predetermined audio. The user provides predetermined audio data and acquires transcription data of the audio through the information processing apparatus according to the embodiment. For example, in the case of a piano, keystroke information (data with keys and time as axes or the like) is acquired as the transcription data. Note that the audio data is a sound wave signal and is data in which amplitude and time are axes. The user terminal 10 may be any apparatus as long as the processing in the embodiment can be realized. The user terminal 10 may be an apparatus such as a smartphone, a tablet terminal, a notebook PC, a desktop PC, a mobile phone, or a PDA. Fig. 3 illustrates an example where the user terminal 10 is a smartphone.

[0056] The user terminal 10 is a smart device such as a smartphone or tablet and is a mobile terminal apparatus capable of communicating with any server apparatus via a wireless communication network such as 4G to 5G (Generation) and LTE (Long Term Evolution). The user terminal 10 may include a screen such as a liquid crystal display having a touch panel function and may receive various operations from the user such as a tap operation, a slide operation, and a scroll operation performed with a finger, a stylus, or the like on display data such as content. In Fig. 3, the user terminal 10 is used by the user U1.

[0057] In Fig. 3, the information processing apparatus 100 acquires audio data to be processed (step S1). For example, the information processing apparatus 100 acquires audio data provided by the user U1 via the user terminal 10 as audio data to be processed.

[0058] The information processing apparatus 100 may acquire, as the audio data to be processed, audio data that is input, specified, or selected by the user U1 via a predetermined application or on the Web using the user terminal 10.

[0059] Further, for example, the information processing apparatus 100 may acquire, as the audio data to be processed, audio data played in real time by the user U1 or the like. For example, the information pro-

cessing apparatus 100 may acquire, as the audio data to be processed, audio data played in real time by the user U1 or the like and provided in real time by the user U1 via the user terminal 10. In this manner, the information processing apparatus 100 may acquire audio data that is real data as the audio data to be processed.

[0060] The information processing apparatus 100 inputs the acquired audio data into the trained machine learning model M1 to generate transcription data corresponding to the audio data (step S2). The information processing apparatus 100 provides this as transcription data of the audio desired by the user U1 (step S3).

[0061] The audio data to be processed may be a piece of music (a certain complete piece of music), only a certain portion, section, or track of a piece of music, or only a certain musical instrument part (for example, only a part for playing a piano) of a piece of music and is not particularly limited and may be any audio data.

[0062] By using the information processing according to the embodiment described above, even when the information processing is applied to a domain with poor or unavailable teacher data, which is a set of annotation data (corresponding to MIDI data) corresponding to the audio and audio data, it is possible to automatically generate general-purpose and highly accurate transcription data. That is, by using the information processing according to the embodiment described above, it is possible to realize an annotation-free AMT.

[0063] Further, by using the "One Shot" audio data, it is possible to generate the synthetic data D1 which is simple and scalable. For example, the synthetic data D1 can be generated by simply cutting and pasting the sample data "One Shot" from among various options to the MIDI data F1.

[0064] Since the "One Shot" audio data may be a real data (for example, data recorded of the sound of an actual performance) or synthetic data (for example, data generated by a synthesizer or the like), the synthetic data D1 may be generated by rendering the "One Shot" real data, the synthetic data D1 may be generated by rendering the "One Shot" synthetic data, or the synthetic data D1 may be generated by rendering a combination of the "One Shot" real data and the "One Shot" synthetic data, thus enabling the synthetic data D1 corresponding to various tones to be generated.

[0065] Further, by using the "One Shot" audio data, it is possible to generate the highly accurate synthetic data D1 that matches music of the Desk Top Music (DTM) music creation era. Since DTM is one form of music that is generally heard frequently in daily life, audio data that can reproduce familiar sound can be created by matching the DTM. Thus, it is possible to promote the improvement of the accuracy from the practical viewpoint.

Variation 1 of Information Processing: Release Time

[0066] In the embodiment described above, the re-

lease time may be taken into consideration. The release time relates to, for example, a reverberation sound remaining after the hand lifts off the key in the case of a piano. The MIDI data F1 corresponds to how long it takes for the sound to attenuate after the offset end and differs depending on the musical instrument or the like. Such release time considerations are generally important in AMT.

[0067] The information processing apparatus 100 may perform rendering taking into consideration the release time of audio in rendering. For example, the information processing apparatus 100 may determine the release time of the audio and apply the release time to the MIDI data F1 to perform rendering. For example, the information processing apparatus 100 may randomly determine a release time within a certain range based on information relating to the musical instrument type included in the MIDI data F1, and may perform rendering, with the sample data being faded out by the length of the release time at the time of rendering.

Variation 2 of Information Processing: Synthesizer

[0068] In the embodiment described above, an example is described where the information processing apparatus 100 trains the machine learning model M1 using the synthetic data D1 generated by rendering the MIDI data F1 and the sample data F2 and performing the additive synthesizing of the sample data F2 according to the MIDI data F1, but the information processing apparatus 100 may train the machine learning model M1 using the synthetic data D1 generated by a synthesizer or the like.

[0069] For example, the information processing apparatus 100 may train the machine learning model M1 using the synthetic data D1 generated from the MIDI data F1 by a synthesizer or the like. As described above, the information processing apparatus 100 may train the machine learning model M1 using the synthetic data D1 generated from the MIDI data F1 by a synthesizer or the like instead of rendering the "One Shot" audio data included in the sample data F2.

Experimental Results

[0070] Hereinafter, experimental results acquired using a proposed method according to the embodiment will be described. Fig. 4 is a diagram illustrating an example of experimental results acquired using a proposed method according to the embodiment. A method T1 corresponds to results acquired by a known method, and a method T3 corresponds to results acquired by a proposed method. A method T2 corresponds to results of a method that is different from the proposed method in that annotation is used, but is similar to the proposed method in other respects.

[0071] First, items in the horizontal row of the table illustrated in Fig. 4 will be described. The "a" of "Real

Data" indicates whether or not learning was performed using real data and indicates that learning was performed using real data for the items in the columns with a check mark.

[0072] The "b" of "Real Data" indicates whether or not learning has been performed using real data and annotation data and indicates that learning has been performed using the real data and the annotation data for the items in the columns with the check mark.

[0073] The "c" of "Real Data" indicates whether or not learning has been performed using more data and indicates that learning has been performed using more data for the items in the columns with the check mark. That is, "b" and "c" of "Real Data" indicate whether or not learning has been performed using a set of audio data and annotation data as teacher data.

[0074] In addition, in the horizontal row items, "Guitarset" corresponds to a result for testing whether or not a real sound of a guitar can be transcribed, "Maestro" corresponds to a result for testing whether or not a real sound of a piano can be transcribed, "Monila" corresponds to a result for testing whether or not a real sound of singing can be transcribed, "Phenicx" corresponds to a result for testing whether or not a real performance of an orchestra can be transcribed, and "Slakh" corresponds to a result for testing whether or not various real synthesized sounds can be transcribed.

[0075] Further, "Fno", "F", and "Acc" are each an index indicating the accuracy of note estimation. Among these, "Fno" is generally the most important index and is an index based on whether or not the onset is within 50 ms. Further, "F" is an index based on whether or not the offset is within 50 ms. Further, "Acc" is an index based on the frame level.

[0076] Based on "Fno", "F", and "Acc", the start and end of a note, the accuracy of pitch, and the like are determined. Further, the values of "Fno", "F", and "Acc" are calculated based on the output MIDI data and the real MIDI data.

[0077] Next, the items in the columns of the table illustrated in Fig. 4 will be described. "Bittner et al.", "Wu et al.", "Gardner et al.", and "Simon et al." correspond to results acquired by a known method and correspond to the above-mentioned NPL 1 to NPL 4, respectively. In addition, "Synthetic-DC", "Synthetic-DA", "Synthetic-L", "Synthetic-M", and "Synthetic-S" correspond to results acquired by the proposed method according to the embodiment.

[0078] Among these, "Synthetic-DC" corresponds to a result of learning performed so that the synthetic data D1 and the real data D2 cannot be distinguished. That is, the result corresponds to the result of "Domain Confusion". Note that "DC" in "Synthetic-DC" is an abbreviation of "Domain Confusion". "Synthetic-DC" corresponds to a result of "Domain Confusion" performed using the synthetic data D1 generated using the "One Shot" audio data.

[0079] In a similar manner, "Synthetic-DA" corresponds to a result of learning performed so that the synthetic data D1 is mistaken for the real data D2. That is, the result corresponds to the result of "Domain Adaptation". Note that "DA" in "Synthetic-DA" is an abbreviation of "Domain Adaptation". "Synthetic-DA" corresponds to a result of "Domain Adaptation" performed using the synthetic data D1 generated using the "One Shot" audio data.

[0080] On the other hand, "Synthetic-L", "Synthetic-M", and "Synthetic-S" correspond to results of learning performed using the synthetic data D1 generated using the "One Shot" audio data without performing "Domain Adaptation" or "Domain Confusion" (that is, without performing the second learning according to the above-described embodiment).

[0081] Note that "L", "M", and "S" in "Synthetic-L", "Synthetic-M", and "Synthetic-S" are abbreviations of "Large", "Medium", and "Small", and the data sizes of the synthetic data D1 are different from each other (the data size of "Synthetic-L" is the largest, and the data size of "Synthetic-S" is the smallest). For example, the data size of the MIDI data F1 and the data size of the sample data F2, which are used for the synthetic data D1, are different.

[0082] "Synthetic-L" is the synthetic data D1 generated in accordance with more complex MIDI data F1 by using a larger number of sample data F2 (for example, a larger number of sample data F2 in terms of the number and variety of "One Shot"). Since the data size of "Synthetic-L" is the largest, a more detailed sound can be expressed, and a tone closer to reality can be acquired. Thus, in the experimental results, the accuracy tends to be the highest among "Synthetic-L", "Synthetic-M", and "Synthetic-S".

[0083] In addition, "Real Mix" corresponds to a result when learning is performed by mixing real audio data with a model similar to that of the proposed method according to the embodiment. Specifically, it corresponds to results of learning and evaluation using all of "Guitarset", "Maestro", "Monila", "Phenicx", and "Slakh" described above. For example, in the case of "Guitarset", the evaluation result corresponds to a result of performing learning using all of "Guitarset", "Maestro", "Monila", "Phenicx", and "Slakh" and evaluating "Guitarset".

[0084] In addition, "Real Omit" corresponds to a result of performing learning and evaluation by excluding the data set of the target domain in "Real Mix". For example, in the case of "Guitarset", the evaluation result corresponds to a result of performing learning using all of "Maestro", "Monila", "Phenicx", and "Slakh" except for "Guitarset" and evaluating "Guitarset". "Real Mix" and "Real Omit" are different from the proposed method according to the embodiment in that learning is performed using annotation data. "Real Mix" and "Real Omit" are results of evaluation for comparison with the proposed method according to the embodiment.

1-3. Configuration of User Terminal according to Embodiment

[0085] Next, a configuration of the user terminal 10 according to the embodiment will be described using Fig. 5. Fig. 5 is a diagram illustrating a configuration example of the user terminal 10 according to the embodiment. As illustrated in Fig. 5, the user terminal includes a communication unit 11, an input unit 12, an output unit 13, and a control unit 14.

[0086] The communication unit 11 is implemented by, for example, a Network Interface Card (NIC). The communication unit 11 is connected with a predetermined network N in a wired or wireless manner and exchanges information with the information processing apparatus 100 and the like via the predetermined network N.

[0087] The input unit 12 receives various types of operations from the user. In Fig. 3, various operations from the user U1 are received. For example, the input unit 12 may receive various operations from the user via the display surface by a touch panel function. The input unit 12 may receive various operations from a button provided in the user terminal 10 or from a keyboard or a mouse connected to the user terminal 10.

[0088] The output unit 13 is a display screen of a tablet terminal or the like realized by, for example, a liquid crystal display, an organic Electro-Luminescence (EL) display, or the like, and is a display apparatus for displaying various types of information. For example, the output unit 13 displays information transmitted from the information processing apparatus 100.

[0089] The control unit 14 is a controller and is implemented by, for example, a Central Processing Unit (CPU), a Micro Processing Unit (MPU), or the like executing various types of programs stored in the storage apparatus in the user terminal 10 using a Random Access Memory (RAM) or the like as a work area. For example, the various types of programs include a program of an application installed in the user terminal 10. For example, the various types of programs include an application program for converting audio played by the user into audio data. The control unit 14 may be implemented by an integrated circuit such as an Application Specific Integrated Circuit (ASIC) and a Field Programmable Gate Array (FPGA).

[0090] As illustrated in Fig. 5, the control unit 14 includes a receiving unit 141 and a transmitting unit 142 and realizes or executes the operations of information processing described below.

[0091] The receiving unit 141 receives various types of information from another information processing apparatus such as the information processing apparatus 100. For example, the receiving unit 141 receives the generated transcription data corresponding to the predetermined audio data. For example, the receiving unit 141 receives the generated transcription data corresponding to the audio data provided by the user.

[0092] Further, for example, the receiving unit 141 receives the generated transcription data corresponding to the audio data input, designated, or selected by the user. Further, for example, the receiving unit 141 receives the generated transcription data corresponding to the audio data of the audio being played by the user. Further, for example, the receiving unit 141 receives the generated transcription data corresponding to the audio data of the audio being played by something/someone other than the user.

[0093] The transmitting unit 142 transmits various types of information to another information processing apparatus such as the information processing apparatus 100. For example, the transmitting unit 142 transmits information indicating that the user desires the transcription data corresponding to the predetermined audio. For example, the transmitting unit 142 transmits predetermined audio data. For example, the transmitting unit 142 transmits audio data input, designated, or selected by the user.

[0094] For example, the transmitting unit 142 transmits information indicating that the user desires the transcription data corresponding to the audio being played by the user. For example, the transmitting unit 142 transmits audio data being played by the user. For example, the transmitting unit 142 transmits information indicating that the user desires the transcription data corresponding to the audio being played by something/someone other than the user. For example, the transmitting unit 142 transmits audio data being played by something/someone other than the user.

1-4. Configuration of Information Processing Apparatus according to Embodiment

[0095] Next, a configuration of the information processing apparatus 100 according to the embodiment will be described using Fig. 6. Fig. 6 is a diagram illustrating a configuration example of the information processing apparatus 100 according to the embodiment. As illustrated in Fig. 6, the information processing apparatus 100 includes a communication unit 110, a storage unit 120, and a control unit 130. The information processing apparatus 100 may include an input unit (for example, a keyboard, a mouse, and the like) for receiving various types of operations from the administrator of the information processing apparatus 100, a display unit (for example, a liquid crystal display or the like) for displaying various types of information, and the like.

[0096] The communication unit 110 is implemented by, for example, a NIC. The communication unit 110 is connected with the network N in a wired or wireless manner and exchanges information with the user terminal 10 and the like via the network N.

[0097] The storage unit 120 is implemented by, for example, a semiconductor memory element such as a RAM or a flash memory or a storage device such as a hard disk or an optical disk. As illustrated in Fig. 6, the

storage unit 120 includes a MIDI data storage unit 121, a sample data storage unit 122, and a real data storage unit 123.

[0098] The MIDI data storage unit 121 stores information relating to the MIDI data (corresponding to the MIDI data F1 in Fig. 1). Here, Fig. 7 illustrates an example of the MIDI data storage unit 121 according to the embodiment. As illustrated in Fig. 7, the MIDI data storage unit 121 includes items such as "MIDI data ID" and "MIDI data".

[0099] "MIDI data ID" indicates identification information for identifying the MIDI data. "MIDI data" indicates MIDI data. For example, information included in the MIDI data is stored in "MIDI data". For example, in the example of Fig. 2, information such as "note N1 (pitch: G3#, time: 2.5 seconds to 3.4 seconds), note N2 (pitch: B3, time: 2.5 seconds to 3.5 seconds), note N3 (pitch: D4, time: 2.5 seconds to 3.6 seconds), and note N4 (pitch: D5, time: 2.5 seconds to 3.0 seconds)" is stored.

[0100] The sample data storage unit 122 stores information relating to sample data (corresponding to the "One Shot" audio data in Fig. 1). Here, Fig. 8 illustrates an example of the sample data storage unit 122 according to the embodiment. As illustrated in Fig. 8, the sample data storage unit 122 includes items such as "Sample data ID" and "Sample data".

[0101] "Sample data ID" indicates identification information for identifying the sample data. "Sample data" indicates sample data. In the example illustrated in Fig. 8, conceptual information such as "sample data #1" or "sample data #2" is stored in "Sample data", but in practice, sound wave data (for example, data in which amplitude and time are axes) is stored. For example, using the example of Fig. 1, the sound wave data corresponding to the sample data F21 may be stored, and the sound wave data corresponding to the main audio Z1 of the sample data F21 may be stored.

[0102] The real data storage unit 123 stores information relating to real data for training the machine learning model M1 (corresponding to the real data D2 in Fig. 1). Here, Fig. 9 illustrates an example of the real data storage unit 123 according to the embodiment. As illustrated in Fig. 9, the real data storage unit 123 includes items such as "Real data ID" and "Real data".

[0103] "Real data ID" indicates identification information for identifying the real data. "Real data" indicates real data. In the example illustrated in Fig. 9, conceptual information such as "Real data #1" or "Real data #2" is stored in "Real data", but in practice, sound wave data is stored. For example, in the example of Fig. 1, the sound wave data corresponding to the real data D2 may be stored, or the feature amount Xr acquired by encoding may be stored.

[0104] The control unit 130 is a controller and is implemented by, for example, the CPU, MPU, or the like executing various programs stored in the storage apparatus inside the information processing apparatus 100

using the RAM as a work area. Also, the control unit 130, for example, is implemented by an integrated circuit such as an ASIC or an FPGA.

[0105] As illustrated in Fig. 6, the control unit 130 includes an acquisition unit 131, a rendering unit 132, an editing unit 133, a conversion unit 134, a training unit 135, and a generation unit 136 and realizes or executes the action of information processing described below. The internal configuration of the control unit 130 is not limited to the configuration illustrated in Fig. 6 and may be another configuration as long as the configuration can execute the information processing to be described later.

[0106] The acquisition unit 131 acquires various types of information. For example, the acquisition unit 131 acquires MIDI data (corresponding to the MIDI data F1 in Fig. 1) and sample data (corresponding to the sample data F2 in Fig. 1).

[0107] For example, the acquisition unit 131 acquires the synthetic data (corresponding to the synthetic data D1 in Fig. 1) and the real data (corresponding to the real data D2 in Fig. 1). For example, the acquisition unit 131 acquires real data according to the actual use of the AMT.

[0108] In addition, for example, the acquisition unit 131 acquires audio data to be processed. Further, for example, the acquisition unit 131 acquires information indicating that the user desires transcription data corresponding to the audio data to be processed.

[0109] The rendering unit 132 renders the MIDI data (corresponding to the MIDI data F1 in Fig. 1) and the sample data (corresponding to the sample data F2 in Fig. 1) acquired by the acquisition unit 131. Specifically, the rendering unit 132 renders the MIDI data and the "One Shot" audio data.

[0110] In the example of Fig. 1, the rendering unit 132 renders the MIDI data F1 and the information included in the sample data F2 corresponding to the note information included in the MIDI data F1. Specifically, the rendering unit 132 renders the sample data from sample data F21 to sample data F23 included in the sample data F2. That is, the rendering unit 132 renders the sample data obtained by combining the synthesized main audio Z1 and sub audio Z2.

[0111] Further, for example, the rendering unit 132 renders the sample data such as the main audio Z1 of the sample data F21 and the sub audio Z2 of the sample data F21. That is, the rendering unit 132 renders the sample data of either the main audio Z1 or the sub audio Z2.

[0112] For example, the rendering unit 132 renders the MIDI data and the sample data to generate synthetic data (corresponding to the synthetic data D1 in Fig. 1), which is audio data corresponding to the MIDI data. In this manner, the rendering unit 132 generates a set of MIDI data and audio data corresponding to the MIDI data.

[0113] The editing unit 133 edits the synthetic data

generated by the rendering unit 132 (corresponding to the editing processing R1 in Fig. 1). Specifically, the editing unit 133 executes preprocessing for training the machine learning model M1. For example, the editing unit 133 compresses the dynamics of the synthetic data so that the volume is constant. In addition, for example, the editing unit 133 applies limiting to the dynamics of the synthetic data based on a threshold. The editing unit 133 then uses the edited synthetic data as a synthetic data (corresponding to the synthetic data D1 in Fig. 1) for training the machine learning model M1.

[0114] The conversion unit 134 encodes the synthetic data (corresponding to the synthetic data D1 in Fig. 1) acquired by the acquisition unit 131, generated by the rendering unit 132, or edited by the editing unit 133 and converts it into a feature amount (corresponding to the feature amount Xs in Fig. 1).

[0115] The conversion unit 134 encodes the real data (corresponding to the real data D2 in Fig. 1) acquired by the acquisition unit 131 and converts it into a feature amount (corresponding to the feature amount Xr in Fig. 1).

[0116] The training unit 135 trains the machine learning model M1. Specifically, the training unit 135 trains the machine learning model M1 based on the two types of feature amounts converted by the conversion unit 134. For example, the training unit 135 trains the machine learning model M1 based on synthetic data (corresponding to the synthetic data D1 in Fig. 1) acquired by synthesizing a plurality of pieces of audio sample data (corresponding to the "One Shot" audio data in Fig. 1). The training unit 135 executes the following two types of learning processing.

[0117] For example, as one learning processing executed by the training unit 135, the training unit 135 trains the machine learning model M1 based on synthetic data (corresponding to the synthetic data D1 in Fig. 1) acquired by synthesizing a plurality of corresponding pieces of audio sample data (corresponding to the "One Shot" audio data in Fig. 1) according to the MIDI data (corresponding to the MIDI data F1 in Fig. 1).

[0118] For example, the training unit 135 trains the machine learning model M1 based on synthetic data (corresponding to the synthetic data D1 in Fig. 1) acquired by synthesizing sample data (corresponding to from sample data F21 to sample data F23 and the like in Fig. 1) with each piece of sample data including at least one type of main audio (corresponding to the main audio Z1 in Fig. 1) and sub audio (corresponding to the sub audio Z2 in Fig. 1).

[0119] Further, for example, the training unit 135 trains the machine learning model M1 based on synthetic data (corresponding to the synthetic data D1 in Fig. 1) acquired by synthesizing sample data (corresponding to sample data such as the main audio Z1 of the sample data F21 and the sub audio Z2 of the sample data F21 in Fig. 1) with each piece of sample data including either sample data of the main audio type

(corresponding to the main audio Z1 in Fig. 1) or sample data of the sub audio type (corresponding to the sub audio Z2 in Fig. 1).

[0120] In the example of Fig. 1, for example, the training unit 135 trains the machine learning model M1 based on the synthetic data D1 acquired by synthesizing the sample data of the main audio Z1 of the sample data F21, the sample data of the sub audio Z2 of the sample data F22, and the like.

[0121] For example, the training unit 135 trains the machine learning model M1 based on synthetic data (corresponding to the synthetic data D1 in Fig. 1) acquired by synthesizing the sample data weighted per type. For example, the training unit 135 trains the machine learning model M1 based on synthetic data (corresponding to the synthetic data D1 in Fig. 1) acquired by synthesizing sample data weighted per type according to whether the data is main audio (corresponding to the main audio Z1 in Fig. 1) or sub audio (corresponding to the sub audio Z2 in Fig. 1).

[0122] For example, the training unit 135 trains the machine learning model M1 based on synthetic data (corresponding to the synthetic data D1 in Fig. 1) acquired by synthesizing sample data of different musical instruments per type. For example, the training unit 135 trains the machine learning model M1 based on synthetic data (corresponding to the synthetic data D1 in Fig. 1) acquired by synthesizing sample data of main audio (corresponding to the main audio Z1 in Fig. 1) and the sub audio (corresponding to the sub audio Z2 in Fig. 1) of different musical instruments per type.

[0123] For example, the training unit 135 trains the machine learning model M1 based on post-application synthetic data (corresponding to the synthetic data D1 in Fig. 1) acquired by applying the predetermined editing processing. For example, the training unit 135 trains the machine learning model M1 based on synthetic data (corresponding to the synthetic data D1 in Fig. 1) acquired by editing executed by the editing unit 133.

[0124] As the second learning processing executed by the training unit 135, for example, the training unit 135 trains the machine learning model M1 to reduce the divergence between the synthetic data (corresponding to the synthetic data D1 in Fig. 1) and the real data (corresponding to the real data D2 in Fig. 1).

[0125] For example, the training unit 135 trains the machine learning model M1 including at least an encoder (the encoder 20 in Fig. 1), a decoder (the decoder 30 in Fig. 1), and a discriminator (the discriminator 40 in Fig. 1).

[0126] In addition, for example, the training unit 135 trains the machine learning model M1 using a predetermined loss function (BCE Loss or the like). For example, the training unit 135 trains the machine learning model M1 using "Classification Loss" (the "Classification Loss" L1 in Fig. 1).

[0127] For example, the training unit 135 trains a discriminator (the discriminator 40 in Fig. 1) so that a dis-

criminator (the discriminator 40 in Fig. 1) can discriminate between synthetic data (corresponding to the synthetic data D1 in Fig. 1) and real data (corresponding to the real data D2 in Fig. 1) when a set of synthetic data (corresponding to the synthetic data D1 in Fig. 1) and real data (corresponding to the real data D2 in Fig. 1) is input into an encoder (corresponding to the encoder 20 in Fig. 1).

[0128] In this manner, with "Classification Loss" (the "Classification Loss" L1 in Fig. 1), the training unit 135 trains the machine learning model M1 to maximize the error (loss), the error being the difference between the result acquired from synthetic data (corresponding to the synthetic data D1 in Fig. 1) via an encoder (the encoder 20 in Fig. 1) and the result acquired from real data (corresponding to the real data D2 in Fig. 1) via an encoder (the encoder 20 in Fig. 1).

[0129] The training of the machine learning model M1 using such "Classification Loss" (the "Classification Loss" L1 in Fig. 1) will be described as "first learning" as appropriate below.

[0130] For example, the training unit 135 trains the machine learning model M1 using "Adversarial Loss" (the "Adversarial Loss" L2 in Fig. 1). For example, the training unit 135 trains an encoder (the encoder 20 in Fig. 1) so that a discriminator (the discriminator 40 in Fig. 1) cannot discriminate between synthetic data (corresponding to the synthetic data D1 in Fig. 1) and real data (corresponding to the real data D2 in Fig. 1) when a set of synthetic data (corresponding to the synthetic data D1 in Fig. 1) and real data (corresponding to the real data D2 in Fig. 1) is input into an encoder (corresponding to the encoder 20 in Fig. 1).

[0131] In other words, for example, after the first learning, the training unit 135 trains an encoder (the encoder 20 in Fig. 1) so that it is impossible for a discriminator (the discriminator 40 in Fig. 1) to discriminate between synthetic data (corresponding to the synthetic data D1 in Fig. 1) and real data (corresponding to the real data D2 in Fig. 1) in an adversarial manner to the first learning.

[0132] With such "Adversarial Loss" (the "Adversarial Loss" L2 in Fig. 1), the training unit 135 trains the machine learning model M1 to minimize the error, the error being the difference between the result acquired from synthetic data (corresponding to the synthetic data D1 in Fig. 1) via an encoder (the encoder 20 in Fig. 1) and the result acquired from real data (corresponding to the real data D2 in Fig. 1) via an encoder (the encoder 20 in Fig. 1).

[0133] The training of the machine learning model M1 using such "Adversarial Loss" (the "Adversarial Loss" L2 in Fig. 1) will be described as "second learning" as appropriate below.

[0134] Also, for example, the training unit 135 trains an encoder (the encoder 20 in Fig. 1) so that a discriminator (the discriminator 40 in Fig. 1) mistakes synthetic data (corresponding to the synthetic data D1 in

Fig. 1) for real data (corresponding to the real data D2 in Fig. 1) when a set of synthetic data (corresponding to the synthetic data D1 in Fig. 1) and real data (corresponding to the real data D2 in Fig. 1) is input into an encoder (corresponding to the encoder 20 in Fig. 1).

[0135] In other words, for example, the training unit 135 trains an encoder (the encoder 20 in Fig. 1) so that a discriminator (the discriminator 40 in Fig. 1) mistakes synthetic data (corresponding to the synthetic data D1 in Fig. 1) for real data (corresponding to the real data D2 in Fig. 1) after the first learning in an adversarial manner to the first learning.

[0136] With such "Adversarial Loss" (the "Adversarial Loss" L2 in Fig. 1), the training unit 135 trains the machine learning model M1 to maximize the error in the reverse direction, the error being the difference between the result acquired from synthetic data (corresponding to the synthetic data D1 in Fig. 1) via an encoder (the encoder 20 in Fig. 1) and the result acquired from real data (corresponding to the real data D2 in Fig. 1) via an encoder (the encoder 20 in Fig. 1).

[0137] The training of the machine learning model M1 using such "Adversarial Loss" (the "Adversarial Loss" L2 in Fig. 1) will be described as "third learning" as appropriate below.

[0138] The generation unit 136 inputs the audio data to be processed acquired by the acquisition unit 131 to the machine learning model M1 to generate corresponding transcription data. For example, the generation unit 136 inputs the audio data to be processed acquired by the acquisition unit 131 to the trained machine learning model M1 trained using audio sample data to generate corresponding transcription data.

[0139] Next, the processing of each unit provided in the information processing apparatus 100 described above will be described in detail according to the flow with reference to Figs. 10 and 11. Fig. 10 is a flowchart illustrating the flow of learning processing in the control unit 130.

[0140] The information processing apparatus 100 determines whether or not MIDI data has been acquired (step S11).

[0141] When the information processing apparatus 100 determines that MIDI data has been acquired (YES in step S11), the information processing apparatus 100 generates synthetic data by rendering the sample data according to the acquired MIDI data (step S12).

[0142] On the other hand, when the information processing apparatus 100 determines that MIDI data has not been acquired (NO in step S11), the information processing apparatus 100 waits until MIDI data is acquired.

[0143] The information processing apparatus 100 acquires the synthetic data and the real data (step S13). For example, the information processing apparatus 100 acquires the synthetic data generated in step S12. The information processing apparatus 100 then

converts the acquired synthetic and real data into feature amounts (step S14).

[0144] The information processing apparatus 100 inputs the feature amount of the synthetic data and the feature amount of the real data to the encoder and further performs the first learning via the discriminator (step S15).

[0145] The information processing apparatus 100 inputs the feature amount of the synthetic data and the feature amount of the real data to the encoder and further performs the second learning or the third learning via the discriminator (step S16).

[0146] Subsequently, the information processing apparatus 100 executes application processing using the trained machine learning model M1. This processing will now be described with reference to Fig. 11. Fig. 11 is a flowchart illustrating the flow of application processing in the control unit 130.

[0147] The information processing apparatus 100 determines whether or not the audio data to be processed has been acquired (step S21).

[0148] When the information processing apparatus 100 determines that the audio data to be processed has been acquired (YES in step S21), the information processing apparatus 100 generates transcription data corresponding to the audio data to be processed by inputting the acquired audio data to be processed into a predetermined machine learning model (corresponding to the trained machine learning model M1) (step S22).

[0149] On the other hand, when the information processing apparatus 100 determines that the audio data to be processed has not been acquired (NO in step S21), the information processing apparatus 100 waits until the audio data to be processed is acquired.

[0150] Then, the information processing apparatus 100 provides the generated transcription data (step S23).

2. Other Embodiments

[0151] The processing according to the above-described embodiment may be executed in various different modes other than the above-described embodiment.

[0152] Further, among the processing operations described in the above-described embodiment, all or some of the processing operations described as being automatically performed may be manually performed, or all or some of the processing operations described as being manually performed may be automatically performed by a known method. In addition, the processing procedures, specific names, and information including various data and parameters described in the above document and drawings can be arbitrarily changed unless otherwise specified. For example, the various types of information illustrated in the drawings are not limited to the illustrated information.

[0153] The components of the devices illustrated in the drawings are functionally conceptual and are not necessarily physically configured as illustrated in the drawings. That is, specific embodiments of distribution and integration of the respective devices are not limited to those illustrated in the drawings, and all or part of each device can be functionally or physically distributed or integrated in any unit according to various loads, usage conditions, or the like.

[0154] The embodiments and the modification examples described above can be combined as appropriate as long as the processing contents do not contradict each other.

[0155] In addition, the effects described in the present specification are merely examples and are not limited, and other effects may be obtained.

3. Effects of Information Processing System according to Present Disclosure

[0156] As described above, an information processing system according to the present disclosure (the information processing system 1 in the embodiment) includes an acquisition unit (the acquisition unit 131 in the embodiment) and a generation unit (the generation unit 136 in the embodiment). The acquisition unit acquires audio data to be processed. The generation unit inputs the audio data acquired by the acquisition unit into the trained machine learning model trained using audio sample data to generate corresponding transcription data.

[0157] In this manner, the information processing system according to the present disclosure can automatically generate general-purpose and highly accurate transcription data corresponding to the audio data to be processed.

[0158] The information processing system further includes a training unit (the training unit 135 in the embodiment). The training unit trains the machine learning model based on synthetic data acquired by synthesizing a plurality of pieces of audio sample data.

[0159] In this manner, the information processing system can enable training of a machine learning model for automatic generation of transcription data using a plurality of pieces of audio sample data.

[0160] The training unit trains the machine learning model based on synthetic data acquired by synthesizing a corresponding plurality of pieces of audio sample data according to MIDI data.

[0161] In this manner, the information processing system can enable training of a machine learning model for automatic generation of transcription data by using MIDI data and generating synthetic data from a plurality of pieces of audio sample data.

[0162] The training unit trains the machine learning model so that the divergence between the synthetic data and the real data is reduced.

[0163] In this manner, the information processing

system can automatically generate general-purpose and highly accurate transcription data when the audio data to be processed is either synthetic data or real data.

5 **[0164]** The training unit trains a machine learning model including at least an encoder, a decoder, and a discriminator.

10 **[0165]** In this manner, the information processing system can train a machine learning model that can reduce the divergence between synthetic data and real data due to including a discriminator in addition to an encoder and a decoder.

15 **[0166]** The training unit performs first learning for training the discriminator so that the discriminator can discriminate between the synthetic data and the real data when the set of the synthetic data and the real data is input to the encoder.

20 **[0167]** In this manner, the information processing system can appropriately identify whether the input information input to the encoder is synthetic data or real data.

25 **[0168]** The training unit performs, after the first learning, second learning for training the encoder so that the discriminator cannot discriminate between the synthetic data and the real data in an adversarial manner to the first learning.

30 **[0169]** In this manner, since the encoder is trained so that the discriminator cannot discriminate between the synthetic data and the real data, the information processing system can automatically generate general-purpose and highly accurate transcription data when the audio data to be processed is either synthetic data or real data.

35 **[0170]** The training unit performs, after the first learning, third learning for training the encoder so that the discriminator mistakes the synthetic data for the real data in an adversarial manner to the first learning.

40 **[0171]** In this manner, since the encoder is trained so that the discriminator mistakes the synthetic data for the real data, the information processing system can automatically generate general-purpose and highly accurate transcription data when the audio data to be processed is either synthetic data or real data. Also, in a case where machine learning is performed to train the encoder so that the discriminator cannot discriminate between the synthetic data and the real data, the information processing system can automatically generate general-purpose and highly accurate transcription data.

50 **[0172]** The training unit trains the machine learning model using a loss function such as Binary Cross Entropy Loss (BCE Loss).

55 **[0173]** In this manner, the information processing system can promote improvement in the accuracy of discriminating between synthetic data and real data using a loss function such as BCE Loss and can automatically generate general-purpose and highly accurate transcription data when the audio data to be pro-

cessed is either synthetic data or real data.

[0174] The training unit trains the machine learning model based on synthetic data acquired by synthesizing sample data with each piece of sample data including at least a type of main audio and sub audio.

[0175] In this manner, the information processing system generates synthetic data from the sample data including the plurality of types of audio data and thus can automatically generate general-purpose and highly accurate transcription data even when the audio data to be processed is complex music including a plurality of tones.

[0176] Also, the training unit trains the machine learning model based on synthetic data acquired by synthesizing the sample data weighted per type.

[0177] In this manner, the information processing system generates synthetic data from the sample data including the plurality of types of audio data weighted per type. This enables training of the machine learning model personalized for each user or each piece of audio data to be processed.

[0178] Also, the training unit trains the machine learning model based on synthetic data acquired by synthesizing the sample data of different musical instruments per type.

[0179] In this manner, the information processing system generates synthetic data from the sample data including the plurality of types of audio data played using different musical instruments per type and thus can automatically generate general-purpose and highly accurate transcription data powered by AI.

[0180] The training unit trains the machine learning model based on the post-application synthetic data acquired by applying the editing processing based on the synthetic data.

[0181] In this manner, the information processing system can use synthetic data close to more realistic audio data as the teacher data of the machine learning model, and thus, when the audio data to be processed is the real data, it is possible to automatically generate general-purpose and highly accurate transcription data to a greater degree.

[0182] The acquisition unit acquires audio data, which is real data, as audio data to be processed.

[0183] In this manner, the information processing system can automatically generate general-purpose and highly accurate transcription data corresponding to audio data that is real data.

4. Hardware Configuration

[0184] An information device such as the information processing apparatus 100 and the user terminal 10 according to each of the above-described embodiments is implemented by a computer 1000 with a configuration as illustrated in Fig. 12, for example. The information processing apparatus 100 according to the embodiment will be described below using an example. Fig.

12 is a diagram of a hardware configuration illustrating an example of the computer 1000 that realizes a function of the information processing apparatus 100. The computer 1000 includes a CPU 1100, a RAM 1200, a read only memory (ROM) 1300, a hard disk drive (HDD) 1400, a communication interface 1500, and an input/output interface 1600. The units of the computer 1000 are connected together via a bus 1050.

[0185] The CPU 1100 operates based on a program stored in the ROM 1300 or the HDD 1400, and controls each unit. For example, the CPU 1100 deploys the program stored in the ROM 1300 or the HDD 1400 into the RAM 1200, and executes processing corresponding to the various programs.

[0186] The ROM 1300 stores a boot program such as a basic input output system (BIOS) executed by the CPU 1100 at the time of activation of the computer 1000, a program depending on hardware of the computer 1000, and the like.

[0187] The HDD 1400 is a computer-readable recording medium that non-transitorily records a program to be executed by the CPU 1100, data to be used by the program, and the like. Specifically, the HDD 1400 is a recording medium that records the conversion program according to the present disclosure, which is an example of program data 1450.

[0188] The communication interface 1500 is an interface for connecting the computer 1000 to an external network 1550 (e.g., the Internet). For example, the CPU 1100 receives data from other devices or transmits data generated by the CPU 1100 to other devices, via the communication interface 1500.

[0189] The input/output interface 1600 is an interface for connecting the input/output device 1650 and the computer 1000. For example, the CPU 1100 receives data from an input device such as a keyboard or a mouse via the input/output interface 1600. The CPU 1100 also transmits data to an output device such as a display, a speaker, or a printer via the input/output interface 1600. The input/output interface 1600 may function as a media interface that reads a program or the like recorded in a predetermined recording medium (media). The medium is, for example, an optical recording medium such as a digital versatile disc (DVD) or a phase change rewritable disk (PD), a magneto-optical recording medium such as a magneto-optical disk (MO), a tape medium, a magnetic recording medium, a semiconductor memory, or the like.

[0190] For example, when the computer 1000 functions as the information processing apparatus 100 according to the embodiment, the CPU 1100 of the computer 1000 executes the information processing program loaded onto the RAM 1200 to implement the functions of the control unit 130 and the like. The HDD 1400 stores the information processing program according to the present disclosure and the data in the storage unit 120. Note that the CPU 1100 reads the program data 1450 from the HDD 1400 and executes the program,

but in another example, the CPU 1100 may acquire these programs from another device via the external network 1550.

[0191] Note that the present technology can also adopt configurations such as the following.

(1) An information processing system including:

an acquisition unit configured to acquire audio data to be processed; and

a generation unit configured to generate corresponding transcription data by inputting the audio data acquired by the acquisition unit into a trained machine learning model trained using sample data of audio.

(2) The information processing system according to (1), further including

a training unit configured to train the machine learning model, based on synthetic data acquired by synthesizing pieces of the sample data of a plurality of pieces of audio.

(3) The information processing system according to (2), in which

the training unit trains the machine learning model, based on synthetic data acquired by synthesizing corresponding pieces of the sample data of the plurality of pieces of audio, according to MIDI data.

(4) The information processing system according to (2), in which

the training unit trains the machine learning model such that divergence between the synthetic data and real data is reduced.

(5) The information processing system according to (4), in which

the training unit trains the machine learning model including at least an encoder, a decoder, and a discriminator.

(6) The information processing system according to (5), in which

the training unit performs first learning in which the discriminator is trained such that the discriminator becomes able to discriminate between the synthetic data and the real data when a set of the synthetic data and the real data is input to the encoder.

(7) The information processing system according to (6), in which

the training unit performs, after the first learning, second learning in which the encoder is trained adversarially to the first learning such that the discriminator becomes unable to discriminate between the synthetic data and the real data.

(8) The information processing system according to (6), in which

the training unit performs, after the first learning, third learning in which the encoder is trained adversarially to the first learning such that the discriminator mistakes the synthetic data for the real data.

(9) The information processing system according to any one of (4) to (8), in which

the training unit trains the machine learning model using a loss function of Binary Cross Entropy Loss (BCE Loss).

(10) The information processing system according to any one of (2) to (9), in which

the training unit trains the machine learning model, based on synthetic data acquired by synthesizing pieces of sample data, each piece of the sample data including at least types of main audio and sub audio.

(11) The information processing system according to (10), in which

the training unit trains the machine learning model, based on synthetic data acquired by synthesizing the pieces of sample data weighted for each of the types.

(12) The information processing system according to (10) or (11), in which

the training unit trains the machine learning model, based on synthetic data acquired by synthesizing the pieces of sample data with different musical instruments for each of the types.

(13) The information processing system according to any one of (2) to (12), in which

the training unit
trains the machine learning model, based on
post-application synthetic data acquired by ap-
plying editing processing based on the syn-
thetic data.

5

(14) The information processing system according
to any one of (1) to (13), in which

the acquisition unit
acquires audio data that is real data as the
audio data to be processed.

10

(15) An information processing method including:

15

an information processing apparatus
acquiring, by an information processing appa-
ratus, audio data to be processed; and
generating, by the information processing ap-
paratus, corresponding transcription data by
inputting the acquired audio data into a trained
machine learning model trained using sample
data of audio.

20

(16) A program for causing a computer to function
as:

25

an acquisition unit configured to acquire audio
data to be processed; and
a generation unit configured to generate corre-
sponding transcription data by inputting the
audio data acquired by the acquisition unit into
a trained machine learning model trained using
sample data of audio.

30

35

Reference Signs List

[0192]

1 Information processing system
10 User terminal
11 Communication unit
12 Input unit
13 Output unit
14 Control unit
20 Encoder
30 Decoder
40 Discriminator
100 Information processing apparatus
110 Communication unit
120 Storage unit
121 MIDI data storage unit
122 Sample data storage unit
123 Real data storage unit
130 Control unit
131 Acquisition unit
132 Rendering unit
133 Editing unit

40

45

50

55

134 Conversion unit
135 Training unit
136 Generation unit
141 Receiving unit
142 Transmitting unit
N Network

Claims

1. An information processing system comprising:

an acquisition unit configured to acquire audio
data to be processed; and
a generation unit configured to generate corre-
sponding transcription data by inputting the
audio data acquired by the acquisition unit into
a trained machine learning model trained using
sample data of audio.

2. The information processing system according to
claim 1, further comprising:

a training unit configured to train the machine
learning model, based on synthetic data acquired
by synthesizing pieces of the sample data of a
plurality of pieces of audio.

3. The information processing system according to
claim 2, wherein

the training unit
trains the machine learning model, based on
synthetic data acquired by synthesizing corre-
sponding pieces of the sample data of the plur-
ality of pieces of audio, according to MIDI data.

4. The information processing system according to
claim 2, wherein

the training unit
trains the machine learning model such that
divergence between the synthetic data and
real data is reduced.

5. The information processing system according to
claim 4, wherein

the training unit
trains the machine learning model including at
least an encoder, a decoder, and a discrimina-
tor.

6. The information processing system according to
claim 5, wherein

the training unit
performs first learning in which the discrimina-
tor is trained such that the discriminator be-

comes able to discriminate between the synthetic data and the real data when a set of the synthetic data and the real data is input to the encoder.

7. The information processing system according to claim 6, wherein

the training unit performs, after the first learning, second learning in which the encoder is trained adversarially to the first learning such that the discriminator becomes unable to discriminate between the synthetic data and the real data.

8. The information processing system according to claim 6, wherein

the training unit performs, after the first learning, third learning in which the encoder is trained adversarially to the first learning such that the discriminator mistakes the synthetic data for the real data.

9. The information processing system according to claim 4, wherein

the training unit trains the machine learning model using a loss function of Binary Cross Entropy Loss (BCE Loss).

10. The information processing system according to claim 2, wherein

the training unit trains the machine learning model, based on synthetic data acquired by synthesizing pieces of sample data, each piece of the sample data including at least types of main audio and sub audio.

11. The information processing system according to claim 10, wherein

the training unit trains the machine learning model, based on synthetic data acquired by synthesizing the pieces of sample data weighted for each of the types.

12. The information processing system according to claim 10, wherein

the training unit trains the machine learning model, based on synthetic data acquired by synthesizing the pieces of sample data with different musical

instruments for each of the types.

13. The information processing system according to claim 2, wherein

the training unit trains the machine learning model, based on post-application synthetic data acquired by applying editing processing based on the synthetic data.

14. The information processing system according to claim 1, wherein

the acquisition unit acquires audio data that is real data as the audio data to be processed.

15. An information processing method comprising:

acquiring, by an information processing apparatus, audio data to be processed; and generating, by the information processing apparatus, corresponding transcription data by inputting the acquired audio data into a trained machine learning model trained using sample data of audio.

16. An information processing program for causing a computer to function as:

an acquisition unit configured to acquire audio data to be processed; and a generation unit configured to generate corresponding transcription data by inputting the audio data acquired by the acquisition unit into a trained machine learning model trained using sample data of audio.

Fig. 1

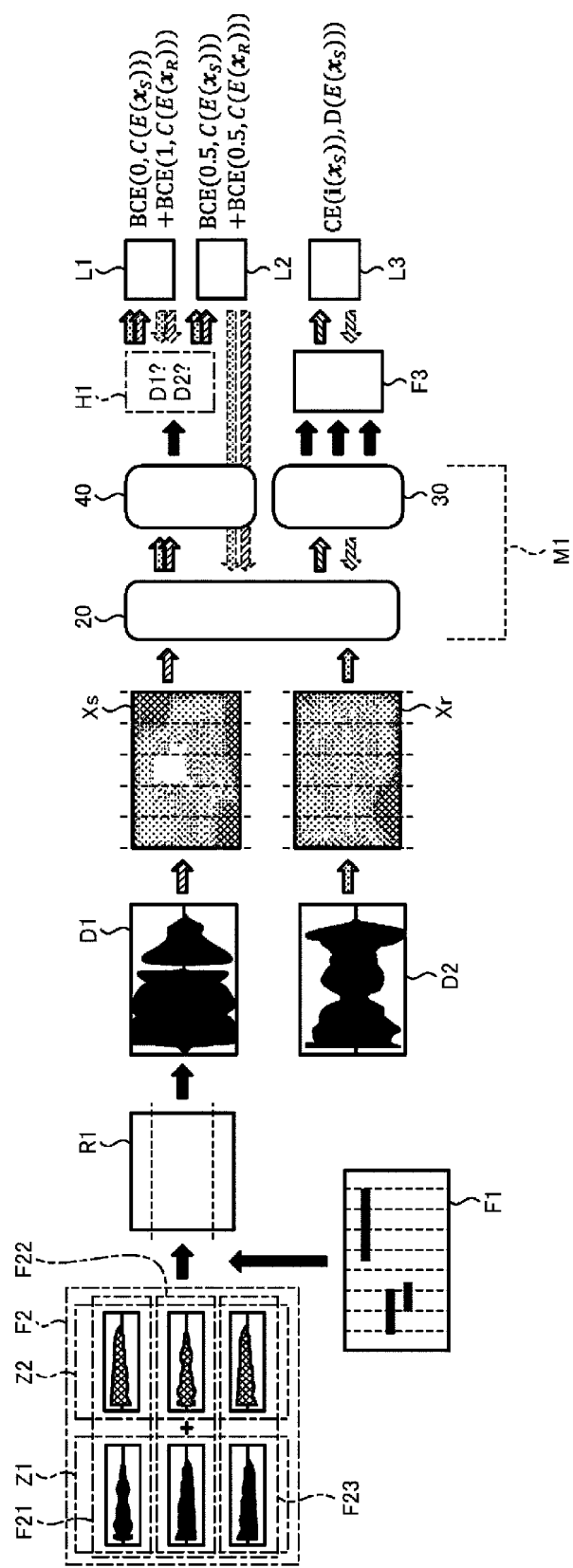


Fig. 2

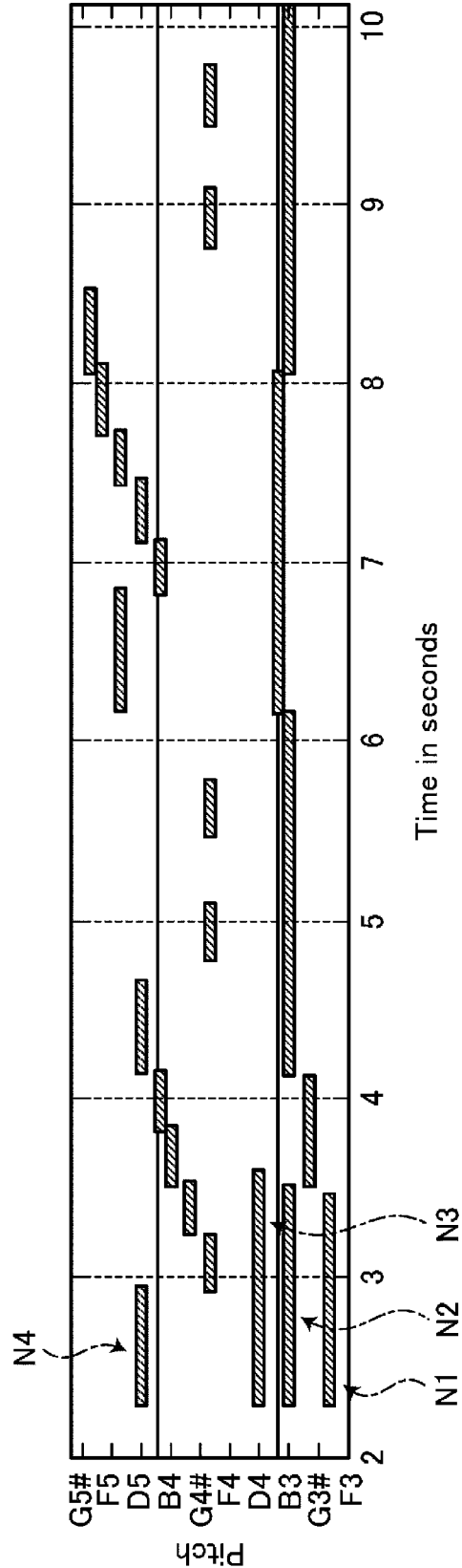


Fig. 3

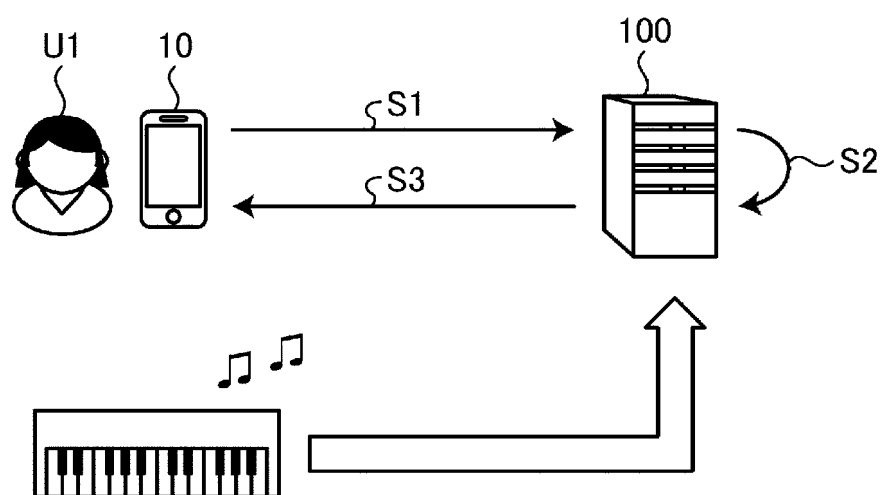


Fig. 4

	Real Data			Our Synth Data		Guitarset		Maestro		Molina		Pheniccx		Slakh						
	a	b	c	Midi	Timbre	Fn	Ac	F	Fn	Ac	F	Fn	Ac	F	Fn	Ac	F			
T1	Bittner et al.	✓	✓	—	—	.79	.70	.56	.71	.38	.11	.55	.60	.38	.51	.50	.36	.43	.40	.23
	Wu et al.	✓	✓	—	—	.59	.43	.27	.30	.39	.07	.31	.48	.11	.12	.13	.05	.23	.40	.07
	Gardner et al.	✓	✓	—	—	.78	—	—	.28	—	—	—	—	—	—	—	—	14*	—	—
	Simon et al.	✓	✓	—	—	.71	—	—	.83	—	—	—	—	—	—	—	—	45*	—	—
T2	Real-Mix	✓	✓	—	—	.80	.76	.58	.90	.57	.54	.78	.78	.58	.66	.66	.40	.79	.79	.60
	Real-Omit	✓	✓	—	—	.78	.64	.43	.51	.47	.22	.26	.66	.13	.61	.51	.17	.35	.42	.13
T3	Synthetic-DC	✓		126K	313K+	.77	.67	.40	.57	.50	.23	.50	.66	.24	.74	.61	.32	.57	.57	.31
	Synthetic-DA	✓		126K	313K+	.64	.60	.27	.50	.49	.20	.47	.67	.23	.64	.63	.28	.54	.54	.28
	Synthetic-L			126K	313K+	.73	.66	.38	.54	.49	.22	.52	.67	.25	.72	.66	.36	.56	.55	.30
	Synthetic-M			126K	792	.71	.65	.35	.55	.47	.21	.47	.68	.20	.70	.60	.25	.57	.57	.31
Synthetic-S				13K	792	.65	.63	.29	.40	.41	.11	.45	.69	.20	.68	.61	.30	.56	.61	.29

T1

T2

T3

Fig . 5

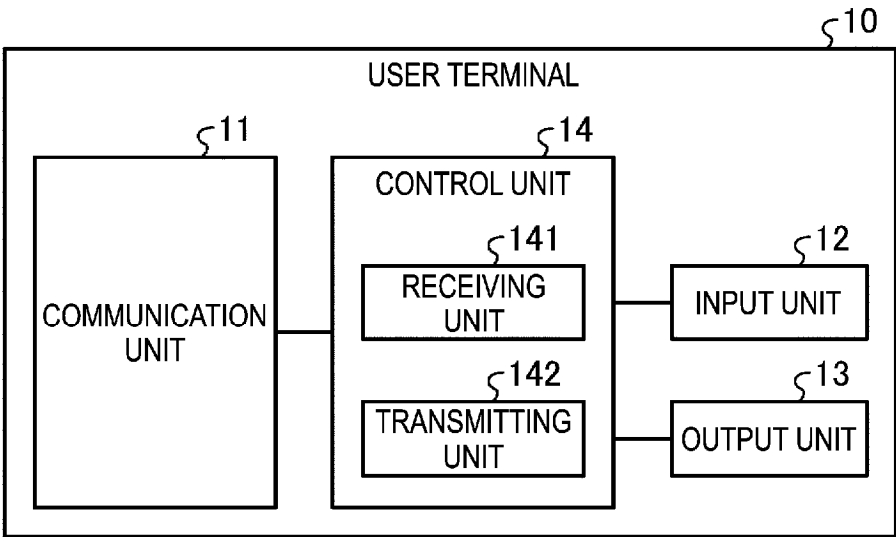


Fig. 6

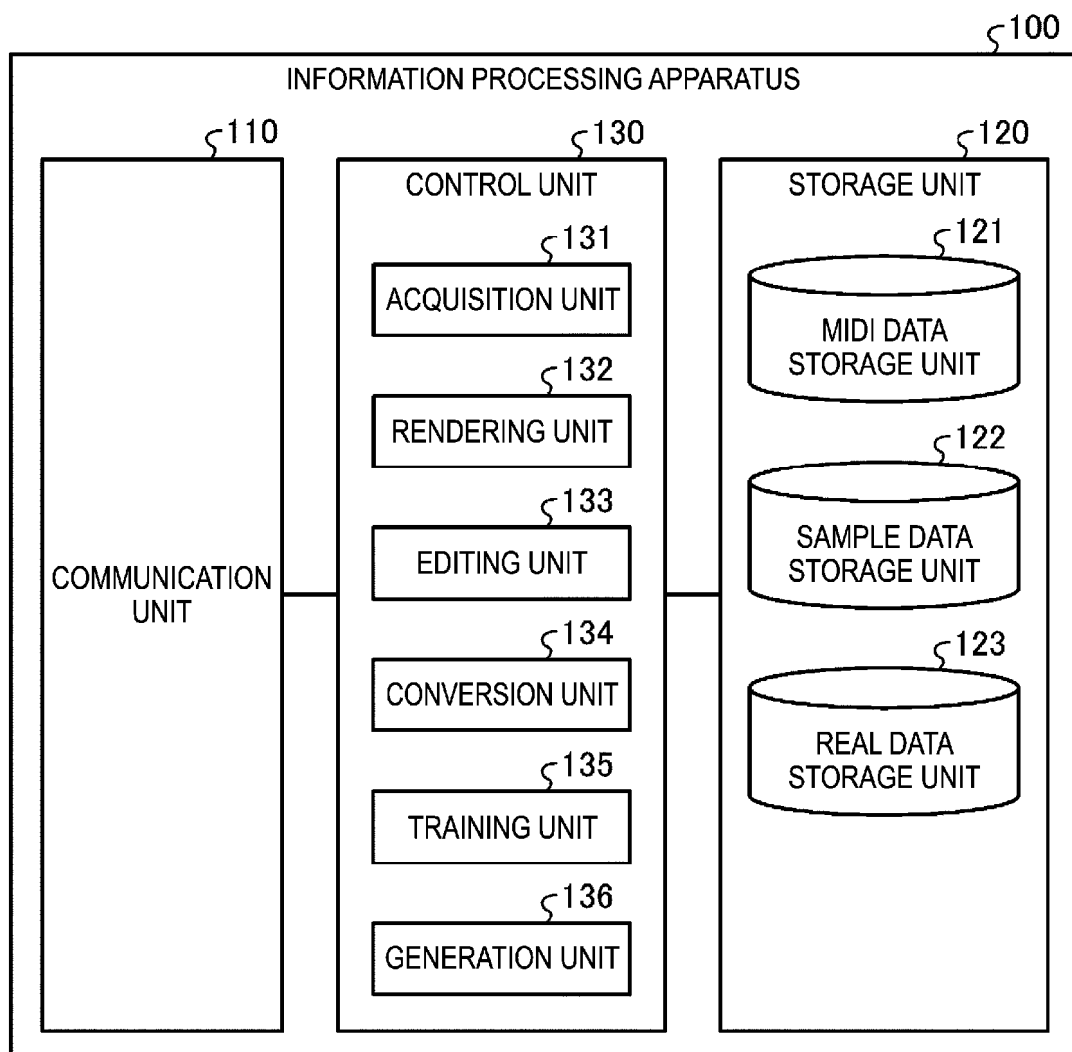


Fig. 7

MIDI DATA ID	MIDI DATA	...
F111	NOTE N1 (PITCH: G3#, TIME: 2.5 SECONDS TO 3.4 SECONDS), NOTE N2 (PITCH: B3, TIME: 2.5 SECONDS TO 3.5 SECONDS), NOTE N3 (PITCH: D4, TIME: 2.5 SECONDS TO 3.6 SECONDS), NOTE N4 (PITCH: D5, TIME: 2.5 SECONDS TO 3.0 SECONDS), ...	...
F112	NOTE N11 (PITCH: G4#, TIME: 3.0 SECONDS TO 3.1 SECONDS), NOTE N12 (PITCH: B4, TIME: 3.1 SECONDS TO 3.2 SECONDS), NOTE N13 (PITCH: D5, TIME: 3.2 SECONDS TO 3.3 SECONDS), NOTE N14 (PITCH: F5, TIME: 3.3 SECONDS TO 3.4 SECONDS), ...	...
...	...	...

ζ121

Fig . 8

§122

SAMPLE DATA ID	SAMPLE DATA	...
F211	SAMPLE DATA #1	...
F212	SAMPLE DATA #2	...
...	...	...

Fig . 9

§123

REAL DATA ID	REAL DATA	...
F311	REAL DATA #1	...
F312	REAL DATA #2	...
...	...	...

Fig. 10

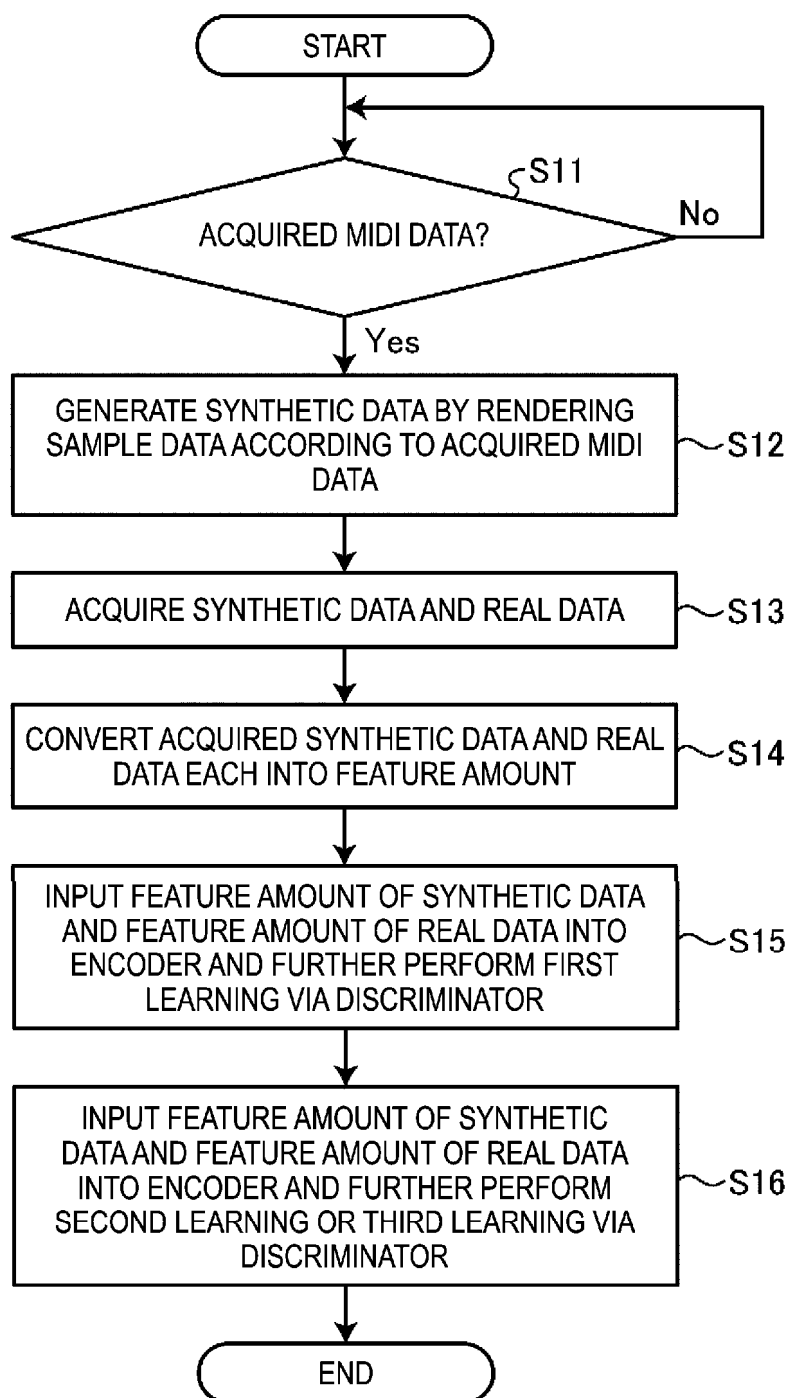


Fig. 11

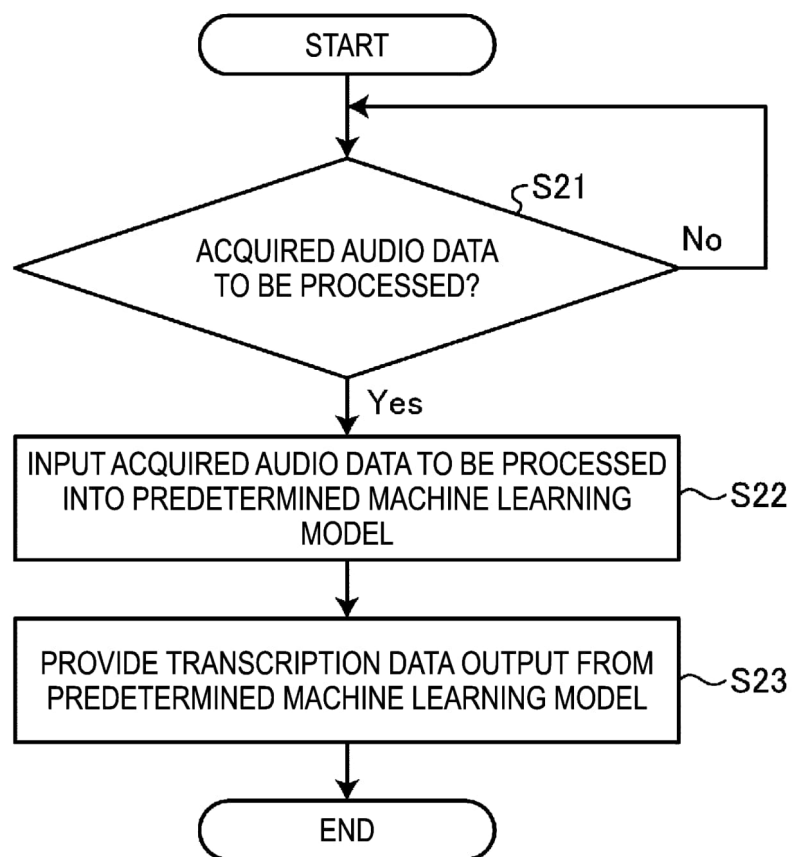
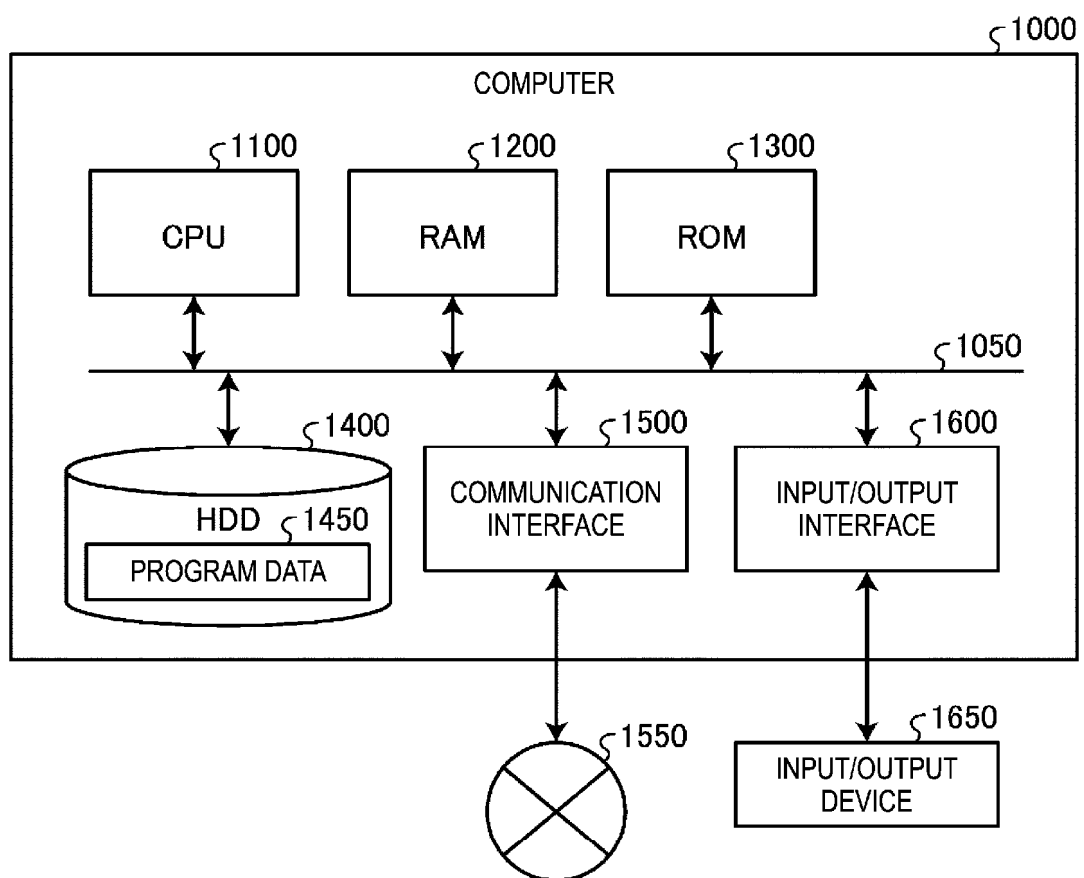


Fig. 12



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2024/026226

A. CLASSIFICATION OF SUBJECT MATTER

G10G 3/04(2006.01)i

FI: G10G3/04

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10G1/00-7/02; G06N3/08; G06N5/04

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Published examined utility model applications of Japan 1922-1996
 Published unexamined utility model applications of Japan 1971-2024
 Registered utility model specifications of Japan 1996-2024
 Published registered utility model applications of Japan 1994-2024

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	KR 10-2021-0123028 A (SK TELECOM CO., LTD.) 13 October 2021 (2021-10-13) paragraphs [0018]-[0023], [0077]-[0091], [0120]-[0130], fig. 1, 6, 10	1-4, 10-12, 14-16
Y	paragraphs [0018]-[0023], [0077]-[0091], [0120]-[0130], fig. 1, 6, 10	5-9, 13
Y	井出優介 ほか, エンコーダ・デコーダ型敵対的生成ネットワークを用いた時系列データの高速異常検知, DEIM2020最終論文集 [online], 2 March 2020, 14-1, [retrieved on 6 September 2024], Internet:<URL: https://db-event.jp.org/deim2020/post/proceedings/#14 >, non-official translation (IDE, Yusuke et al., High-speed anomaly detection in time series data using Encoder-Decoder generative adversarial networks, Proceedings of DEIM 2020) page 1, left column, lines 1-16, page 3, left column, line 38 - right column, line 13, page 4, left column, lines 9-22, fig. 2	5-8
A	entire text, all drawings	1-4, 9-16
Y	JP 2023-81946 A (CASIO COMPUTER CO., LTD.) 13 June 2023 (2023-06-13) paragraphs [0013], [0021]	9, 13
A	entire text, all drawings	1-8, 10-12, 14-16

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“D” document cited by the applicant in the international application

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

06 September 2024

Date of mailing of the international search report

17 September 2024

Name and mailing address of the ISA/JP

Japan Patent Office (ISA/JP)
 3-4-3 Kasumigaseki, Chiyoda-ku, Tokyo 100-8915
 Japan

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/JP2024/026226

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
KR	10-2021-0123028	A	13 October 2021	(Family: none)	
JP	2023-81946	A	13 June 2023	(Family: none)	

Form PCT/ISA/210 (patent family annex) (July 2022)

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Non-patent literature cited in the description

- **J. GARDNER et al.** MT3: multi-task multitrack music transcription. *Proceedings of ICLR*, 2021 [0004]
- **S. IAN et al.** Scaling polyphonic transcription with mixtures of monophonic transcriptions. *Proceedings of ISMIR*, 2022 [0004]
- **R. M. BITTNER et al.** A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation. *Proceedings of ICASSP*, 2022 [0004]
- **Y.-T. WU et al.** Multi-instrument automatic music transcription with self-attention-based instance segmentation. *Proceedings of ICML*, 2022 [0004]