(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 4 181 123 A1

(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
17.05.2023 Bulletin 2023/20

(21) Application number: **22205114.6**

(22) Date of filing: **02.11.2022**

(51) International Patent Classification (IPC):
G10L 25/30 $^{(2013.01)}$ ***G10L 25/63*** $^{(2013.01)}$

(52) Cooperative Patent Classification (CPC):
G10L 25/30; G10L 25/63

(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:
KH MA MD TN

(30) Priority: **11.11.2021 JP 2021183862**

(71) Applicant: **Hitachi, Ltd.**
Tokyo 100-8280 (JP)

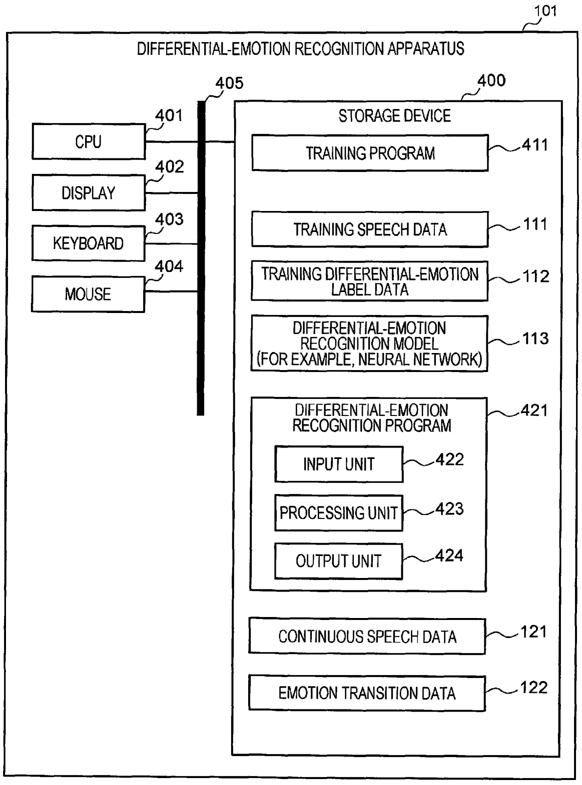
(72) Inventor: **SUMIYOSHI, Takashi**
Tokyo, 100-8280 (JP)

(74) Representative: **MERH-IP Matias Erny Reichl Hoffmann**
Patentanwälte PartG mbB
Paul-Heyse-Strasse 29
80336 München (DE)

(54) EMOTION RECOGNITION SYSTEM AND EMOTION RECOGNITION METHOD

(57) An emotion recognition system includes an input unit (422) configured to input first speech data and second speech data, and a processing unit (423) configured to input the first speech data and the second speech data to a differential-emotion recognition model (113) that infers a differential emotion between two pieces of speech data, and acquire, from the differential-emotion recognition model, differential-emotion information indicating a differential emotion between the first speech data and the second speech data.

FIG. 4



Processed by Luminess, 75001 PARIS (FR)

Description

BACKGROUND OF THE INVENTION

1. Field of the Invention

[0001] The present invention generally relates to a technique for inferring emotions expressed in speech.

2. Description of the Related Art

[0002] Emotion information included in human speech plays an important role in human communication. There is a possibility that whether a purpose of communication has been achieved can be determined by movement of emotions, which creates a demand for analyzing emotions in communication. In business situations such as daily sales activities and reception operations at a call center, it is necessary to analyze emotions in a lot of speech-based communication. Thus, speech emotion recognition by a machine is desired.

[0003] The speech emotion recognition outputs categories of emotions included in input speech or a degree for each emotion category for the speech. As a mechanism thereof, there are a method of performing classification or regression analysis from features of speech signals in accordance with a predetermined rule, and a method of obtaining such a rule by machine learning.

[0004] In recent years, a speech interaction apparatus capable of easily performing emotion recognition according to a user has been disclosed (see JP 2018-132623 A).

SUMMARY OF THE INVENTION

[0005] In a rule-based method, a result that is output on the basis of one piece of speech data depends on a rule or a threshold set by a developer. In a machine learning method, a parameter is determined on the basis of speech data collected for training and a labeling result by a person (labeler) who labels the speech data with emotional impressions received by listening to that speech data as correct values. However, in any case, determination is made depending on subjectivity of the minority such as a developer or a labeler, so that an output of an emotion recognition device may deviate from an actual user impression.

[0006] The present invention has been made in view of the above points, and an object of the present invention is to propose an emotion recognition system and the like capable of appropriately recognizing emotions expressed in speech.

[0007] In order to solve such a problem, the present invention includes an input unit configured to input first speech data and second speech data, and a processing unit configured to input the first speech data and the second speech data to a differential-emotion recognition model that infers a differential emotion between two pieces of speech data, and acquire, from the differential-emotion recognition model, differential-emotion information indicating a differential emotion between the first speech data and the second speech data.

[0008] In the above configuration, for example, since the differential emotion (that is, relative emotion) is inferred from the two pieces of speech data, an emotion recognized by the emotion recognition system can be closer to an actual user impression than an emotion inferred from one piece of speech data (that is, absolute emotion).

[0009] According to the present invention, it is possible to implement an emotion recognition system that recognizes emotions expressed in speech with high accuracy. Problems, configurations, and advantageous effects other than those described above will be clarified by the following description of embodiments.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010]

Fig. 1 is a diagram illustrating an example of a processing flow of a differential-emotion recognition apparatus according to a first embodiment;
Fig. 2 is a diagram illustrating an example of training speech data according to the first embodiment;
Fig. 3 is a diagram illustrating an example of training differential-emotion label data according to the first embodiment;
Fig. 4 is a diagram illustrating an example of a configuration of a differential-emotion recognition apparatus according to the first embodiment;
Fig. 5 is a diagram illustrating an example of a flowchart of a training program according to the first embodiment;
Fig. 6 is a diagram illustrating an example of a flowchart of a differential-emotion recognition program according to the first embodiment;
Fig. 7 is a diagram illustrating an example of emotion transition data according to the first embodiment;

Fig. 8 is a diagram illustrating an example of a user interface according to the first embodiment; and

Fig. 9 is a diagram illustrating an example of a processing flow of a differential-emotion recognition apparatus according to a second embodiment.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

(I) First embodiment

[0011] Hereinafter, embodiments of the present invention will be described in detail. However, the present invention is not limited to the embodiments.

[0012] In a conventional technique, there is a possibility that an unintended result is output because a rule functions incorrectly due to influence of a variable factor of speaker characteristics in input speech, a variable factor of environmental characteristics in input speech, and the like. Recently, the advent of deep learning allows for handling more complicated rules in machine learning, and efforts have been widely made to solve this problem. Although the accuracy has been improved, it is difficult to say that the problem has been sufficiently solved.

[0013] In this regard, an emotion recognition system according to the present embodiment compares speech data within an individual rather than between individuals to recognize speech emotions. The speech emotions are the inside of a person expressed as a voice, including delight, anger, sorrow, and pleasure, negativeness or positiveness, and the like. The emotion recognition system uses an emotion label given by listening to two speech patterns of a certain person (a label of differential assessment (relative assessment) that quantifies emotions grasped from the two speech patterns of the person) rather than an emotion label given by listening to a single speech pattern of a certain person (a label of absolute assessment that quantifies current emotions, for example, the person is happy or sad). Thus, the labeling work is easy for labelers, and the label of the differential assessment is more reliable than the label of the absolute assessment. The two speech patterns used by the emotion recognition system to infer emotions come from the same speaker, but may be discontinuous. However, it is desirable to use speech patterns acquired on the same day and/or at the same place.

[0014] According to the emotion recognition system, it is possible to recognize speech emotions based on a difference between speech patterns of the same person while suppressing the influence of the speaker characteristics and the environmental characteristics as compared with the related art.

[0015] Next, the embodiments of the present invention will be described with reference to the drawings. The following description and drawings are examples for explaining the present invention, and some parts are omitted and simplified as appropriate for the sake of clarity of explanation. The present invention can also be implemented in various other forms. Unless otherwise specified, each constituent element may be singular or plural. In the following description, the same elements are denoted by the same reference numerals in the drawings, and the description thereof will be omitted as appropriate.

[0016] Expressions such as "first", "second", and "third" in the present specification are used to identify constituent elements, and do not necessarily limit the number or order of the constituent elements. In addition, reference numerals for identifying constituent elements are used in each context, and a reference numeral used in one context does not necessarily indicate the same configuration in another context. Furthermore, a constituent element identified by a certain reference numeral is not hindered from concurrently serving a function of a constituent element identified by another reference numeral.

[0017] Fig. 1 is a diagram illustrating an example of a processing flow of a differential-emotion recognition apparatus 101 according to the present embodiment.

[0018] First, in a training phase 110, a user 102 prepares training speech data 111 and training differential-emotion label data 112. Next, the user 102 uses the differential-emotion recognition apparatus 101 to generate a differential-emotion recognition model 113 by training.

[0019] Next, in an inference phase 120, the user 102 inputs continuous speech data 121 to the differential-emotion recognition apparatus 101 and acquires emotion transition data 122.

[0020] Fig. 2 is a diagram illustrating an example (training speech table 200) of the training speech data 111.

[0021] The training speech table 200 stores a plurality of speech waveforms (speech data). Each of the plurality of speech waveforms is given a speech ID and a speaker ID. The speech ID represents a code for uniquely identifying a speech waveform. The speaker ID, which is assigned to a speaker of a speech waveform, represents a code for uniquely identifying a speaker. For that matter, the training speech table 200 stores a plurality of speech waveforms of a plurality of persons.

[0022] Fig. 3 is a diagram illustrating an example (training differential-emotion label table 300) of the training differential-emotion label data 112.

[0023] The training differential-emotion label table 300 stores a plurality of differential emotions. The differential emotion, which is given by a labeler, represents a label that quantifies an emotion of a speech pattern (second speech pattern) having a speech waveform of a second speech ID relative to a speech pattern (first speech pattern) having a speech waveform of a first speech ID. The first speech ID and the second speech ID indicate speech waveforms of the corresponding speech IDs of the training speech data 111.

[0024] It is assumed that the speech waveforms of the first speech ID and the second speech ID have the same speaker ID. The labeler gives an emotion label (of a difference value) between the two inputs. Note that a label of an absolute value of emotion for one speech pattern may be given (held in the training speech data 111) as in the related art. In training, the difference between the absolute values may be used as the difference value. For example, when absolute values "0.1" and "0.2" indicating emotions for the speech patterns of the speech IDs "1" and "2", respectively, are stored in the training speech table 200, the difference value "0.1" between the absolute value "0.1" of the first speech ID and the absolute value "0.2" of the second speech ID may be calculated as a corresponding differential emotion.

[0025] Moreover, the training differential-emotion label table 300 may store differential emotions by a plurality of labelers. In this case, a statistical value such as an average value of the differential emotions by the plurality of labelers is used in training. Furthermore, there may be multiple emotion categories instead of one. In this case, the training differential-emotion label table 300 stores a vector value as a differential emotion instead of a scalar value.

[0026] Fig. 4 is a diagram illustrating an example of a configuration of the differential-emotion recognition apparatus 101.

[0027] The differential-emotion recognition apparatus 101 includes a storage device 400, a CPU 401, a display 402, a keyboard 403, and a mouse 404 as components, like a configuration of a general personal computer (PC). Each component can transmit and receive data via a bus 405.

[0028] The storage device 400 includes a training program 411 and a differential-emotion recognition program 421 as programs. These programs are read to the CPU 401 by an operating system (OS) (not illustrated) existing in the storage device 400 at start-up to be executed.

[0029] The differential-emotion recognition model 113 is, for example, a neural network in which the number of states of an input layer is "1024" in total of the number of features of a first speech pattern "512" and the number of features of a second speech pattern "512", the number of hidden layers is one and the number of states of the hidden layer is "512", and the number of states of an output layer is "1". For the input {x_i (i = 1 ... 1024)} of the input layer, the value {h_j (j = 1 ... 512)} of the hidden layer is calculated by (Expression 1).

[Mathematical formula 1]

$$h_j = s\left(\sum\nolimits_i \left(W_{ij}^1 x_i + b_i^1\right)\right) \quad \text{... (Expression 1)}$$

[0030] The output y of the output layer represents a difference value of emotion of the second speech pattern relative to the first speech pattern, and is calculated by (Expression 2).

[Mathematical formula 2]

$$y = s\left(\sum\nolimits_i \left(W_i^2 h_i + b_i^2\right)\right) \quad \text{... (Expression 2)}$$

[0031] Here, s represents an activation function, for example, a sigmoid function, W represents a weight, and b represents a bias.

[0032] As a means for obtaining the features from the first speech pattern and the second speech pattern, a statistic or the like for time-series low-level descriptors (LLD), which is described in Document 1 below, can be used.

[0033] Document 1: Suzuki, "Recognition of Emotions Included in Speech", the Journal of the Acoustical Society of Japan, Vol. 71, No. 9 (2015), pp. 484-489

[0034] Note that the present embodiment does not limit the structure of the neural network. Any structure and activation function of a neural network may be used. Furthermore, the differential-emotion recognition model 113 is not limited to a neural network, and may adopt any model.

[0035] Functions of the differential-emotion recognition apparatus 101 (training program 411, differential-emotion recognition program 421, and the like) may be implemented by, for example, the CPU 401 reading a program from the storage device 400 and executing the program (software), may be implemented by hardware such as a dedicated circuit, or may be implemented by a combination of software and hardware. Note that one function of the differential-emotion recognition apparatus 101 may be divided into multiple functions, or multiple functions may be integrated into one function. For example, the differential-emotion recognition program 421 may include an input unit 422, a processing unit 423, and an output unit 424. Moreover, some of the functions of the differential-emotion recognition apparatus 101 may be provided as an alternate function or may be included in another function. Furthermore, some of the functions of the differential-emotion recognition apparatus 101 may be implemented by another computer connectable to the differential-emotion recognition apparatus 101. For example, the training program 411 may be provided in a first PC, and the differential-emotion recognition program 421 may be provided in a second PC.

[0036] Fig. 5 is a diagram illustrating an example of a flowchart of the training program 411.

[0037] First, the training program 411 assigns initial values to the parameters W_{ij}^1, b_i^1, W_i^2, and b_i^2 of the differential-emotion recognition model 113 (S501). The training program 411 gives random values as the initial values for facilitating the training of the neural network.

[0038] Next, the training program 411 reads data from the training speech data 111 and the training differential-emotion label data 112 (S502).

[0039] Next, the training program 411 updates the parameters of the differential-emotion recognition model 113 (S503). As an update method, a neural network back propagation method can be used.

[0040] Next, the training program 411 determines whether the training is converged (S504). The convergence is determined under conditions that the process has been executed a predetermined fixed number of times, a value of an error function has fallen below a predetermined threshold, and the like.

[0041] Fig. 6 is a diagram illustrating an example of a flowchart of the differential-emotion recognition program 421.

[0042] Before the differential-emotion recognition program 421 is executed, the user 102 or the like stores speech to be analyzed in the storage device 400 as the continuous speech data 121.

[0043] First, the differential-emotion recognition program 421 reads one frame of the continuous speech data 121. If all the continuous speech data 121 has been read (S601), the program ends.

[0044] Next, the differential-emotion recognition program 421 determines whether a speech section has been detected (S602). A known method can be used to detect a speech section. In the method, for example, when a certain number of frames with a volume level equal to or higher than a given value continue and then a certain number of frames with a volume level equal to or lower than the given value continue, a group of these frames is regarded as a speech section. When no speech sections have been detected, the process returns to S601. Note that the speech section may be a section (a group of sequential frames) in which a speech pattern is detected, or may be a section including a previous frame and/or a subsequent frame of the section in which the speech pattern is detected.

[0045] Next, the differential-emotion recognition program 421 stores information of the detected speech section in the emotion transition data 122 (S603). Note that, in S603, a speech section ID and speech section data are stored in the emotion transition data 122 and, in S606, an emotion transition is stored.

[0046] Next, the differential-emotion recognition program 421 determines whether a pair of speech sections can be selected (S604). The pair of speech sections to be selected may be, for example, a pair including temporally adjacent speech sections in the emotion transition data 122, for one of which no emotion transitions have been calculated. Here, in order to make calculation of an emotion transition robust, the differential-emotion recognition program 421 may use a plurality of pairs of speech sections within a predetermined time interval as the pair of adjacent speech sections. In addition, the differential-emotion recognition program 421 may perform processing of eliminating a pair including a speech section for which emotion recognition is considered to be difficult, such as a short speech section or a speech section with a low volume level.

[0047] Next, the differential-emotion recognition program 421 inputs a selected pair of speech sections (two pieces of speech data of the speech sections) to the differential-emotion recognition model 113, and obtains a differential emotion output from the differential-emotion recognition model 113 (S605).

[0048] Next, the differential-emotion recognition program 421 calculates an emotion transition on the basis of the differential emotion and stores the emotion transition in the emotion transition data 122 (S606). In order to obtain an emotion transition of a certain speech section, for example, the obtained differential emotion may be added to an emotion transition of the other speech section paired with that speech section for all the pairs, and obtained values may be averaged. The initial speech section may have an emotion transition of an average value "0". Thereafter, the process returns to S601.

[0049] In the configuration illustrated in Fig. 6, the differential emotion is acquired by inputting a pair of speech sections detected from the continuous speech data 121 to the differential-emotion recognition model 113. However, the present invention is not limited to this configuration. For example, the differential-emotion recognition program 421 may be configured to acquire the differential emotion by inputting two pieces of speech data designated by the user to the differential-emotion recognition model 113.

[0050] Fig. 7 is a diagram illustrating an example (emotion transition table 700) of the emotion transition data 122.

[0051] The emotion transition table 700 stores an emotion transition for each speech section. The speech section ID represents a code for uniquely identifying a speech section. The speech section data represents information (for example, time section information) indicating a range of the speech section in the continuous speech data 121. The emotion transition represents an emotion transition value obtained by the differential-emotion recognition program 421 for the speech section.

[0052] Fig. 8 is a diagram illustrating an example of a user interface of the differential-emotion recognition apparatus

101.

[0053] The user 102 obtains, from the display 402, information that a continuous speech file to be input can be selected. When the user 102 operates the keyboard 403 and/or the mouse 404 to press a speech file selection button 801 and selects a speech file stored in the differential-emotion recognition apparatus 101, speech in the speech file is visualized on the display 402 as a waveform 810 and is stored in the storage device 400 as the continuous speech data 121. The user 102 can subsequently press an analysis start button 802 to run the differential-emotion recognition program 421. When the emotion transition data 122 is generated, emotion transition values are visualized on the display 402 as a graph 820.

[0054] Note that the emotion transition values are calculated for respective speech sections, and thus are smoothly connected in the graph 820 to show the emotion transitions in time series. For that matter, data to be visualized is not limited to the emotion transition values, and may be categories of emotion or a degree (differential-emotion value) for each emotion category.

[0055] The configuration of the emotion recognition system described above allows for a user to easily look at transition of emotions in speech of a specific speaker using the difference emotion recognition device based on the model that has learned the difference values of emotional expressions in speech.

(II) Second embodiment

[0056] In the case of the conventional system designed to output absolute emotion assessment values (absolute values of emotion) for input speech, a user might input speech of different persons and use emotion assessment values of the speech for emotional assessments of the persons. In some cases, such usage is inappropriate since an output of an emotion recognition device reflects subjectivity of the minority and is not accurate enough as described above. As a main application, it is often sufficient to show a relative emotion change in speech of the same person by, for example, visualizing rise and fall of emotion expressions. However, a mechanism to limit application thereto is not provided.

[0057] In this regard, a differential-emotion recognition apparatus 901 according to the present embodiment determines whether two pieces of input speech data come from an identical speaker. In the present embodiment, the same configurations as those of the first embodiment are denoted by the same reference signs, and the description thereof will be omitted.

[0058] Fig. 9 is a diagram illustrating an example of a processing flow of the differential-emotion recognition apparatus 901 according to the present embodiment.

[0059] A differential-emotion recognition model 911 according to the present embodiment includes an identical-speaker recognition unit and has, as an output layer, an output z in addition to an output y. The output z, which represents an identical-speaker determination value, is "1" when a first speech pattern and a second speech pattern come from an identical speaker, and "0" when the first speech pattern and the second speech pattern do not come from an identical speaker. The output z is calculated by (Expression 3).

[Mathematical formula 3]

$$z = s\left(\sum_i \left(W_i^3 h_i + b_i^3\right)\right) \quad \text{... (Expression 3)}$$

[0060] Note that the differential-emotion recognition model 911 has parameters W_{ij}^1, b_i^1, W_i^2, b_i^2, W_i^3, and b_i^3.

[0061] In the training program 411, when training is performed using training speech data 111 and training differential-emotion label data 112 similar to those in the first embodiment, a label z is set to "1". In addition, any two pieces of speech data having different speaker IDs are retrieved from the training speech data 111. Then, a label y is set to a random value (value unrelated to a differential emotion of the training differential-emotion label data 112), and the label z is set to "0" to update the parameters.

[0062] In the differential-emotion recognition program 421, when an emotion transition is calculated, speakers are determined not to be identical by an identical-speaker determination value less than a threshold. In a case where a pair of which speakers are determined not to be identical accounts for a certain ratio or more of all the pairs, the emotion transition is an invalid value. When emotion transitions are visualized by the display 402, an indication that the emotion recognition is invalid is displayed for a result of the speech section of the invalid value. For that matter, the differential-emotion recognition program 421 may stop the analysis in a case where an emotion transition is an invalid value.

[0063] Note that the present embodiment is not limited to the configuration in which the differential-emotion recognition model 911 includes the identical-speaker recognition unit. For example, the differential-emotion recognition apparatus 901 may include the differential-emotion recognition model 113 and an identical-speaker recognition unit (for example, an identical-speaker recognition model of a neural network).

[0064] The configuration of the emotion recognition system described above prevents a user from obtaining a result in a case where the user attempts the emotion recognition for speech including speech patterns of different speakers. As a result, it is possible to avoid the usage in which emotion assessment values obtained when speech of different persons is input are regarded as emotional assessments of the persons.

(III) Supplementary notes

[0065] The above-described embodiments include, for example, the following contents.

[0066] In the above embodiments, a case where the present invention is applied to an emotion recognition system has been described. However, the present invention is not only applied to such a system but also can be widely applied to other various systems, apparatuses, methods, and programs.

[0067] In the above embodiments, a process is sometimes described with a "program" as the subject. However, the subject of the process may be a processor unit since the program is executed by the processor unit to perform the defined process appropriately using a storage unit (for example, a memory), an interface unit (for example, a communication port), and/or the like. A process described with a program as the subject may be a process performed by the processor unit or by a device having the processor unit. Furthermore, the processor unit may include a hardware circuit that performs a part or all of the process (for example, a field-programmable gate array (FPGA) or an application specific integrated circuit (ASIC)).

[0068] In the above embodiments, a part or all of the programs may be installed from a program source into an apparatus such as a computer that implements the differential-emotion recognition apparatus. The program source may be, for example, a program distribution server connected via a network or a computer-readable recording medium (for example, a non-transitory recording medium). Furthermore, in the above description, two or more programs may be implemented as one program, or one program may be implemented as two or more programs.

[0069] In the above embodiments, the configuration of each table is an example, and one table may be divided into two or more tables, or two or more tables may be combined into one table in whole or in part.

[0070] In the above embodiments, the screen illustrated and described is an example, and any design may be employed as long as the same information can be received.

[0071] In the above embodiments, the screen illustrated and described is an example, and any design may be employed as long as the same information can be presented.

[0072] In the above embodiments, the case where the average value is used as the statistical value has been described. However, the statistical value is not limited to the average value, and may be another statistical value such as a maximum value, a minimum value, a difference between the maximum value and the minimum value, a mode value, a median value, or a standard deviation.

[0073] In the above embodiments, information is not always output by being displayed on the display. The information may be output as audio output by a speaker, may be output to a file, may be output by being printed on a paper medium or the like by a printing apparatus, may be output by being projected on a screen or the like by a projector, or may be output in another mode.

[0074] In the above description, information such as programs, tables, and files for implementing the functions may be stored in a storage device such as a memory, a hard disk, and a solid state drive (SSD), or in a recording medium such as an IC card, an SD card, and a DVD.

[0075] The above-described embodiments have, for example, the following characteristic configurations.

(1) An emotion recognition system (for example, the differential-emotion recognition apparatus 101, the differential-emotion recognition apparatus 901, a system including the differential-emotion recognition apparatus 101 and a computer capable of communicating with the differential-emotion recognition apparatus 101, or a system including the differential-emotion recognition apparatus 901 and a computer capable of communicating with the differential-emotion recognition apparatus 901) includes: an input unit (for example, the differential-emotion recognition program 421, the input unit 422, or a circuit) configured to input first speech data and second speech data; and a processing unit (for example, the differential-emotion recognition program 421, the processing unit 423, or a circuit) configured to input the first speech data and the second speech data to a differential-emotion recognition model (for example, the differential-emotion recognition model 113 or the differential-emotion recognition model 911) that infers a differential emotion between two pieces of speech data, and acquire, from the differential-emotion recognition model, differential-emotion information indicating a differential emotion between the first speech data and the second speech data. Note that the first speech data and the second speech data may be included in one speech data source (for example, the continuous speech data 121) or may be derived from separate speech data sources.

In the above configuration, for example, since the differential emotion (that is, relative emotion) is inferred from the two pieces of speech data, an emotion recognized by the emotion recognition system can be closer to an actual user impression than an emotion inferred from one piece of speech data (that is, absolute emotion).

(2) The emotion recognition system further includes: a determination unit (for example, the differential-emotion recognition program 421 including the determination unit, the determination unit, or a circuit) configured to input the first speech data and the second speech data to an identical-speaker recognition unit (for example, the differential-emotion recognition model 911 or the identical-speaker recognition unit) that infers whether speakers of two pieces of speech data are identical, acquire, from the identical-speaker recognition unit, determination information indicating that a speaker of the first speech data and a speaker of the second speech data are identical (for example, "0" for indicating they are not identical or "1" for indicating they are identical, or a value from "0" to "1"), and determine whether the speaker of the first speech data and the speaker of the second speech data are identical according to the acquired determination information; and an output unit (for example, the differential-emotion recognition program 421 including the determination unit, the output unit 424, or a circuit) configured to output information according to a result of determination by the determination unit.

According to the above configuration, it is determined whether the speakers of the two pieces of speech data are identical. This allows for avoiding incorrect usage in which, for example, speech of different persons is input for emotional assessments of the persons.

(3) The output unit is configured to, when the determination unit determines that the speaker of the first speech data and the speaker of the second speech data are identical, output the differential-emotion information (for example, the graph 820) acquired by the processing unit, and the output unit is configured to, when the determination unit determines that the speaker of the first speech data and the speaker of the second speech data are not identical, output information that the first speech data and the second speech data do not come from an identical speaker (for example, "A speech section of which speakers are not identical is included. This speech section is hidden in the graph.").

According to the above configuration, for example, a mechanism to reject comparison between speech patterns of different persons can be provided.

(4) The input unit is configured to input continuous speech data (for example, the continuous speech data 121), and the processing unit is configured to detect at least one speech section from the continuous speech data input by the input unit, extract a piece of speech data of each of the at least one speech section from the continuous speech data, select, out of extracted pieces of speech data, a first piece of speech data and a second piece of speech data within a predetermined time interval from the first piece of speech data, input the first piece of speech data and the second piece of speech data to the differential-emotion recognition model, and acquire differential-emotion information indicating a differential emotion between the first piece of speech data and the second piece of speech data (for example, see Fig. 6).

According to the above configuration, when the continuous speech data is input, for example, the differential-emotion information indicating the differential emotion between the two adjacent pieces of speech data is sequentially acquired, so that the user can grasp transition of emotions.

(5) The differential-emotion recognition model is trained by using two pieces of speech data of a same person and differential-emotion information indicating a differential emotion between the two pieces of speech data (see Fig. 5).

In the above configuration, labeling is performed on the two pieces of speech data of the same person. Therefore, for example, a labeler easily estimates emotions, and likelihood of labels depending on subjectivity of the labeler can be reduced.

(6) When the identical-speaker recognition unit is trained by using two pieces of speech data, differential-emotion information indicating a differential emotion between the two pieces of speech data, and information indicating whether speakers of the two pieces of speech data are identical, in a case where the two pieces of speech data come from different persons, training is performed by changing the differential-emotion information indicating the differential emotion between the two pieces of speech data to a value (for example, a random value) unrelated to the differential-emotion information.

According to the above configuration, for example, the differential-emotion recognition model and the identical-speaker recognition unit can be trained by using common data, so that a burden of preparing data used for training can be reduced.

(7) The differential-emotion recognition model is a neural network.

[0076] In the above configuration, since the differential-emotion recognition model is a neural network, it is possible to reduce likelihood of a rule functioning incorrectly due to the influence of variable factors of speaker characteristics such as a voice tendency to be muffled and a high voice, environmental characteristics such as large reverberation, and the like. It is also possible to improve the inference accuracy.

[0077] In addition, the above configurations may be appropriately changed, rearranged, combined, or omitted without departing from the gist of the present invention.

[0078] It should be understood that items included in a list in the format "at least one of A, B, and C" can mean (A), (B), (C), (A and B), (A and C), (B and C), or (A, B, and C). Similarly, items listed in the format "at least one of A, B, or

C;" can mean (A), (B), (C), (A and B), (A and C), (B and C), or (A, B, and C).

Claims

1. An emotion recognition system comprising:

    an input unit configured to input first speech data and second speech data; and
    a processing unit configured to input the first speech data and the second speech data to a differential-emotion recognition model that infers a differential emotion between two pieces of speech data, and acquire, from the differential-emotion recognition model, differential-emotion information indicating a differential emotion between the first speech data and the second speech data.

2. The emotion recognition system according to claim 1, further comprising:

    a determination unit configured to input the first speech data and the second speech data to an identical-speaker recognition unit that infers whether speakers of two pieces of speech data are identical, acquire, from the identical-speaker recognition unit, determination information indicating that a speaker of the first speech data and a speaker of the second speech data are identical, and determine whether the speaker of the first speech data and the speaker of the second speech data are identical according to the acquired determination information; and
    an output unit configured to output information according to a result of determination by the determination unit.

3. The emotion recognition system according to claim 2, wherein the output unit is configured to, when the determination unit determines that the speaker of the first speech data and the speaker of the second speech data are identical, output the differential-emotion information acquired by the processing unit, and the output unit is configured to, when the determination unit determines that the speaker of the first speech data and the speaker of the second speech data are not identical, output information that the first speech data and the second speech data do not come from an identical speaker.

4. The emotion recognition system according to claim 1, wherein

    the input unit is configured to input continuous speech data, and
    the processing unit is configured to detect at least one speech section from the continuous speech data input by the input unit, extract a piece of speech data of each of the at least one speech section from the continuous speech data, select, out of extracted pieces of speech data, a first piece of speech data and a second piece of speech data within a predetermined time interval from the first piece of speech data, input the first piece of speech data and the second piece of speech data to the differential-emotion recognition model, and acquire differential-emotion information indicating a differential emotion between the first piece of speech data and the second piece of speech data.

5. The emotion recognition system according to claim 1, wherein the differential-emotion recognition model is trained by using two pieces of speech data of a same person and differential-emotion information indicating a differential emotion between the two pieces of speech data.

6. The emotion recognition system according to claim 2, wherein when the identical-speaker recognition unit is trained by using two pieces of speech data, differential-emotion information indicating a differential emotion between the two pieces of speech data, and information indicating whether speakers of the two pieces of speech data are identical, in a case where the two pieces of speech data come from different persons, training is performed by changing the differential-emotion information indicating the differential emotion between the two pieces of speech data to a value unrelated to the differential-emotion information.

7. The emotion recognition system according to claim 1, wherein the differential-emotion recognition model is a neural network.

8. An emotion recognition method comprising:

    by an input unit, inputting first speech data and second speech data; and by a processing unit, inputting the first speech data and the second speech data to a differential-emotion recognition model that infers a differential emotion between two pieces of speech data, and acquiring, from the differential-emotion recognition model, differential-emotion information indicating a differential emotion between the first speech data and the second speech data.

9. The emotion recognition method according to claim 8, further comprising:

by a determination unit, inputting the first speech data and the second speech data to an identical-speaker recognition unit that infers whether speakers of two pieces of speech data are identical, acquiring, from the identical-speaker recognition unit, determination information indicating that a speaker of the first speech data and a speaker of the second speech data are identical, and determining whether the speaker of the first speech data and the speaker of the second speech data are identical according to the acquired determination information; and

by an output unit, outputting information according to a result of the determination by the determination unit.

10. The emotion recognition method according to claim 9, wherein when the determination unit determines that the speaker of the first speech data and the speaker of the second speech data are identical, the output unit outputs the differential-emotion information acquired by the processing unit, and when the determination unit determines that the speaker of the first speech data and the speaker of the second speech data are not identical, the output unit outputs information that the first speech data and the second speech data do not come from an identical speaker.

11. The emotion recognition method according to claim 8, wherein

the input unit inputs continuous speech data, and

the processing unit detects at least one speech section from the continuous speech data input by the input unit, extracts a piece of speech data of each of the at least one speech section from the continuous speech data, selects, out of extracted pieces of speech data, a first piece of speech data and a second piece of speech data within a predetermined time interval from the first piece of speech data, inputs the first piece of speech data and the second piece of speech data to the differential-emotion recognition model, and acquires differential-emotion information indicating a differential emotion between the first piece of speech data and the second piece of speech data.

12. The emotion recognition method according to claim 8, wherein the differential-emotion recognition model is trained by using two pieces of speech data of a same person and differential-emotion information indicating a differential emotion between the two pieces of speech data.

13. The emotion recognition method according to claim 9, wherein when the identical-speaker recognition unit is trained by using two pieces of speech data, differential-emotion information indicating a differential emotion between the two pieces of speech data, and information indicating whether speakers of the two pieces of speech data are identical, in a case where the two pieces of speech data come from different persons, training is performed by changing the differential-emotion information indicating the differential emotion between the two pieces of speech data to a value unrelated to the differential-emotion information.

14. The emotion recognition method according to claim 8, wherein the differential-emotion recognition model is a neural network.

FIG. 1

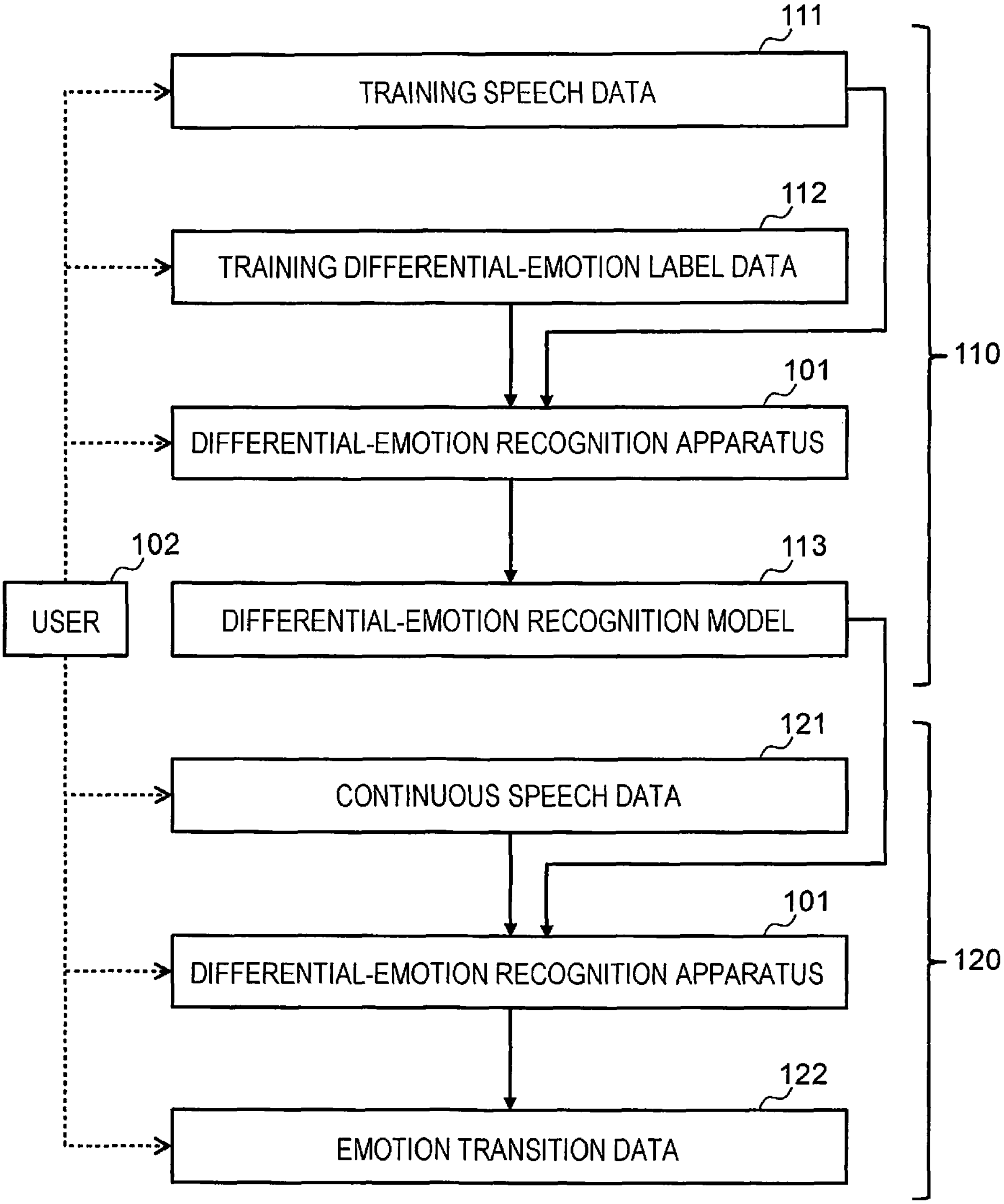

111
TRAINING SPEECH DATA
112
TRAINING DIFFERENTIAL-EMOTION LABEL DATA
101
DIFFERENTIAL-EMOTION RECOGNITION APPARATUS
110
102
USER
113
DIFFERENTIAL-EMOTION RECOGNITION MODEL
121
CONTINUOUS SPEECH DATA
101
DIFFERENTIAL-EMOTION RECOGNITION APPARATUS
120
122
EMOTION TRANSITION DATA

FIG. 2

200

SPEECH ID	SPEAKER ID	SPEECH WAVEFORM
1	1	
2	1	
3	2	

FIG. 3

300

FIRST SPEECH ID	SECOND SPEECH ID	DIFFERENTIAL EMOTION
1	2	0.1
2	1	-1.0

FIG. 4

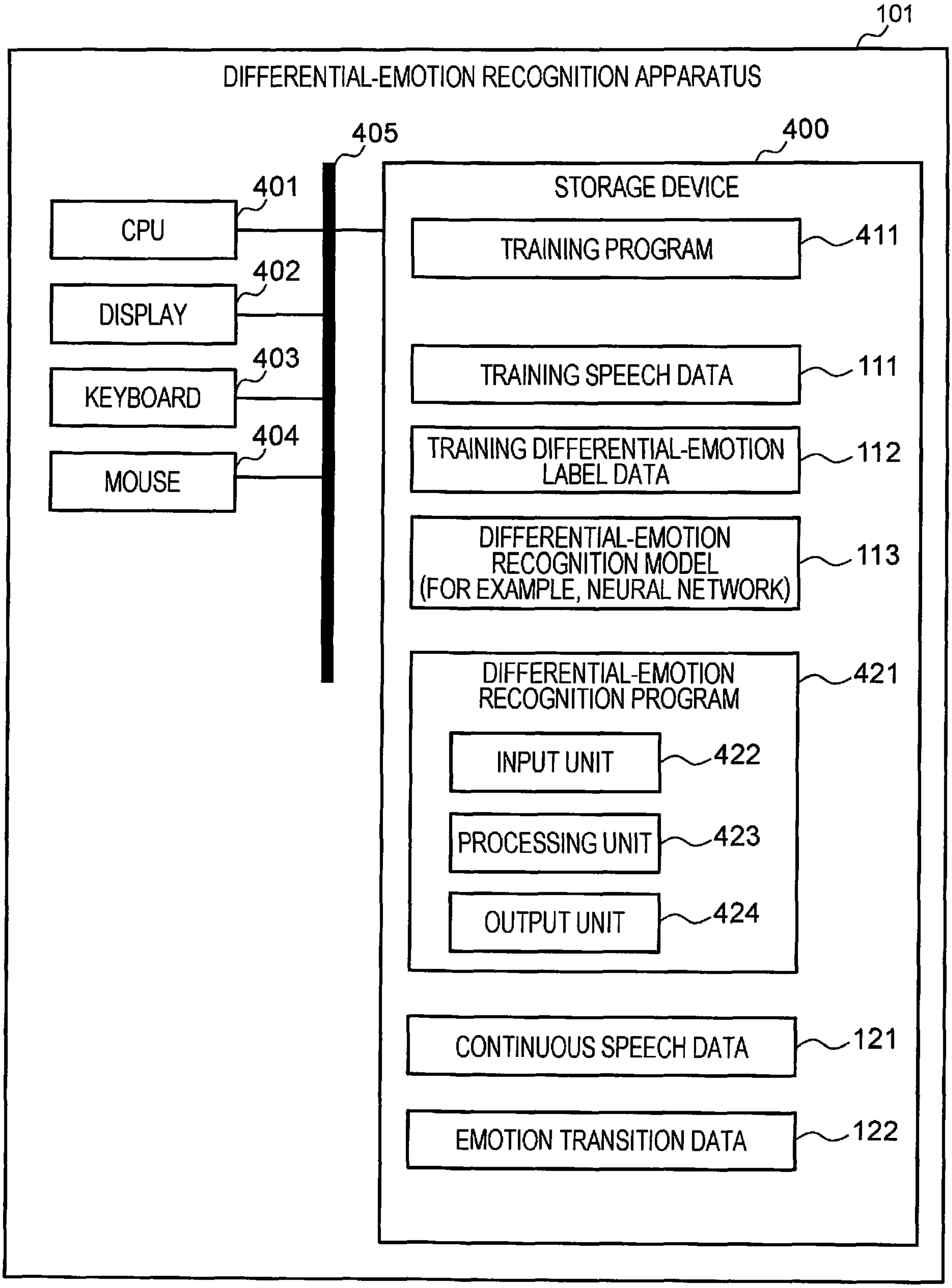

101
DIFFERENTIAL-EMOTION RECOGNITION APPARATUS
405
400
STORAGE DEVICE
401
CPU
402
DISPLAY
403
KEYBOARD
404
MOUSE
TRAINING PROGRAM
411
TRAINING SPEECH DATA
111
TRAINING DIFFERENTIAL-EMOTION LABEL DATA
112
DIFFERENTIAL-EMOTION RECOGNITION MODEL (FOR EXAMPLE, NEURAL NETWORK)
113
DIFFERENTIAL-EMOTION RECOGNITION PROGRAM
421
INPUT UNIT
422
PROCESSING UNIT
423
OUTPUT UNIT
424
CONTINUOUS SPEECH DATA
121
EMOTION TRANSITION DATA
122

FIG. 5

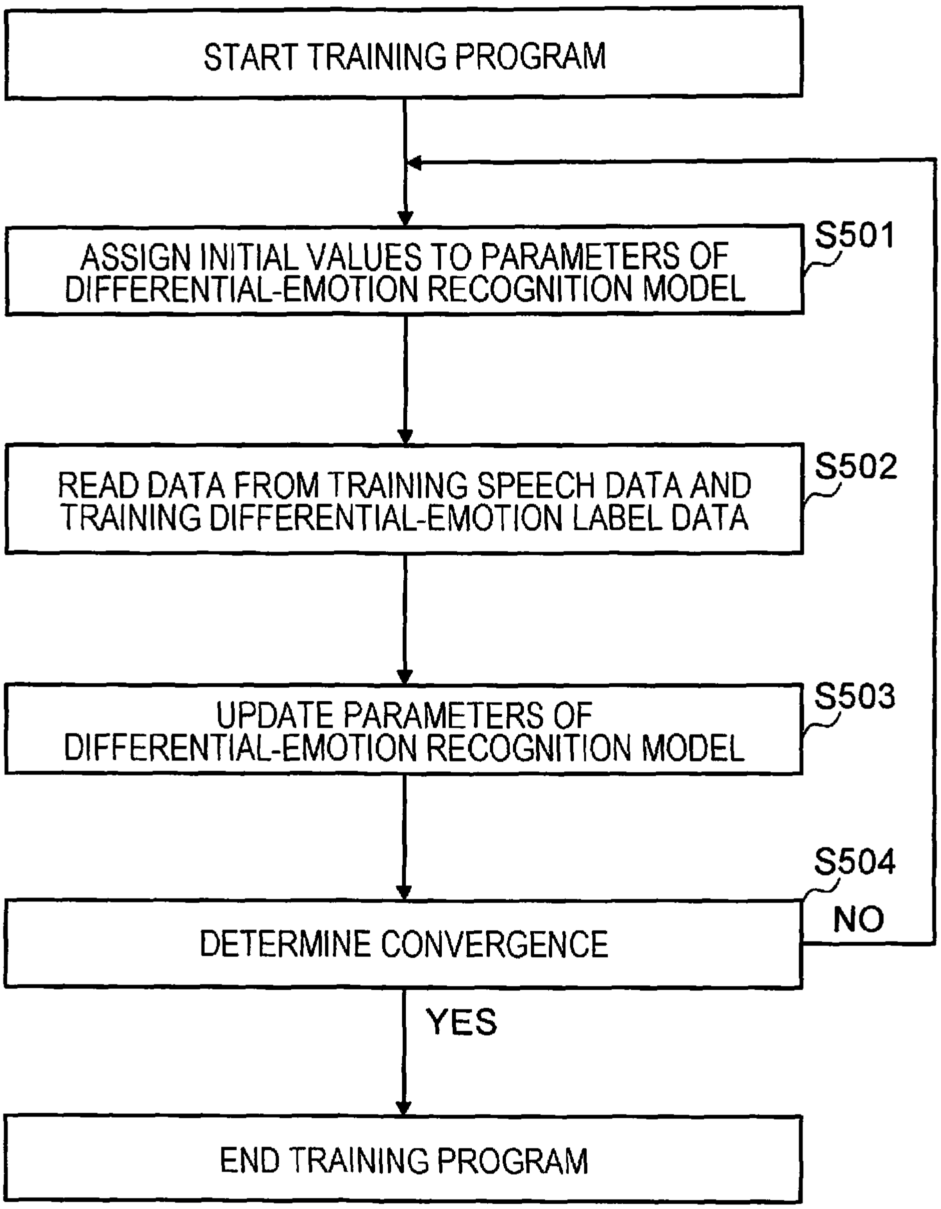

START TRAINING PROGRAM
ASSIGN INITIAL VALUES TO PARAMETERS OF DIFFERENTIAL-EMOTION RECOGNITION MODEL
S501
READ DATA FROM TRAINING SPEECH DATA AND TRAINING DIFFERENTIAL-EMOTION LABEL DATA
S502
UPDATE PARAMETERS OF DIFFERENTIAL-EMOTION RECOGNITION MODEL
S503
S504
NO
DETERMINE CONVERGENCE
YES
END TRAINING PROGRAM

FIG. 6

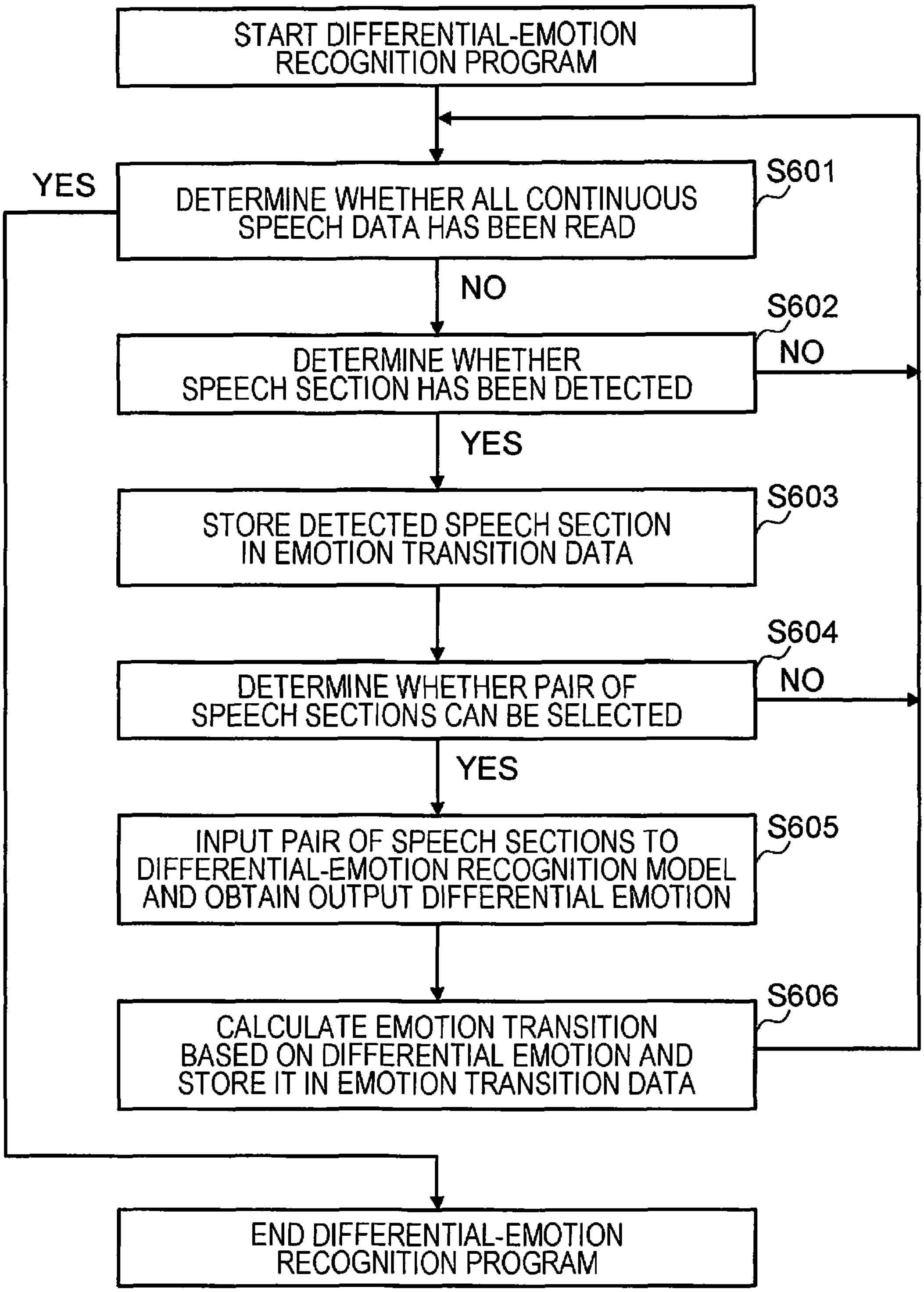

START DIFFERENTIAL-EMOTION RECOGNITION PROGRAM
S601
DETERMINE WHETHER ALL CONTINUOUS SPEECH DATA HAS BEEN READ
YES
NO
S602
DETERMINE WHETHER SPEECH SECTION HAS BEEN DETECTED
NO
YES
S603
STORE DETECTED SPEECH SECTION IN EMOTION TRANSITION DATA
S604
DETERMINE WHETHER PAIR OF SPEECH SECTIONS CAN BE SELECTED
NO
YES
S605
INPUT PAIR OF SPEECH SECTIONS TO DIFFERENTIAL-EMOTION RECOGNITION MODEL AND OBTAIN OUTPUT DIFFERENTIAL EMOTION
S606
CALCULATE EMOTION TRANSITION BASED ON DIFFERENTIAL EMOTION AND STORE IT IN EMOTION TRANSITION DATA
END DIFFERENTIAL-EMOTION RECOGNITION PROGRAM

FIG. 7

700

SPEECH SECTION ID	SPEECH SECTION DATA	EMOTION TRANSITION
1	[0:10—0:13]	0.0
2	[0:20—0:25]	0.2
3	[0:28—0:33]	-0.2

FIG. 8

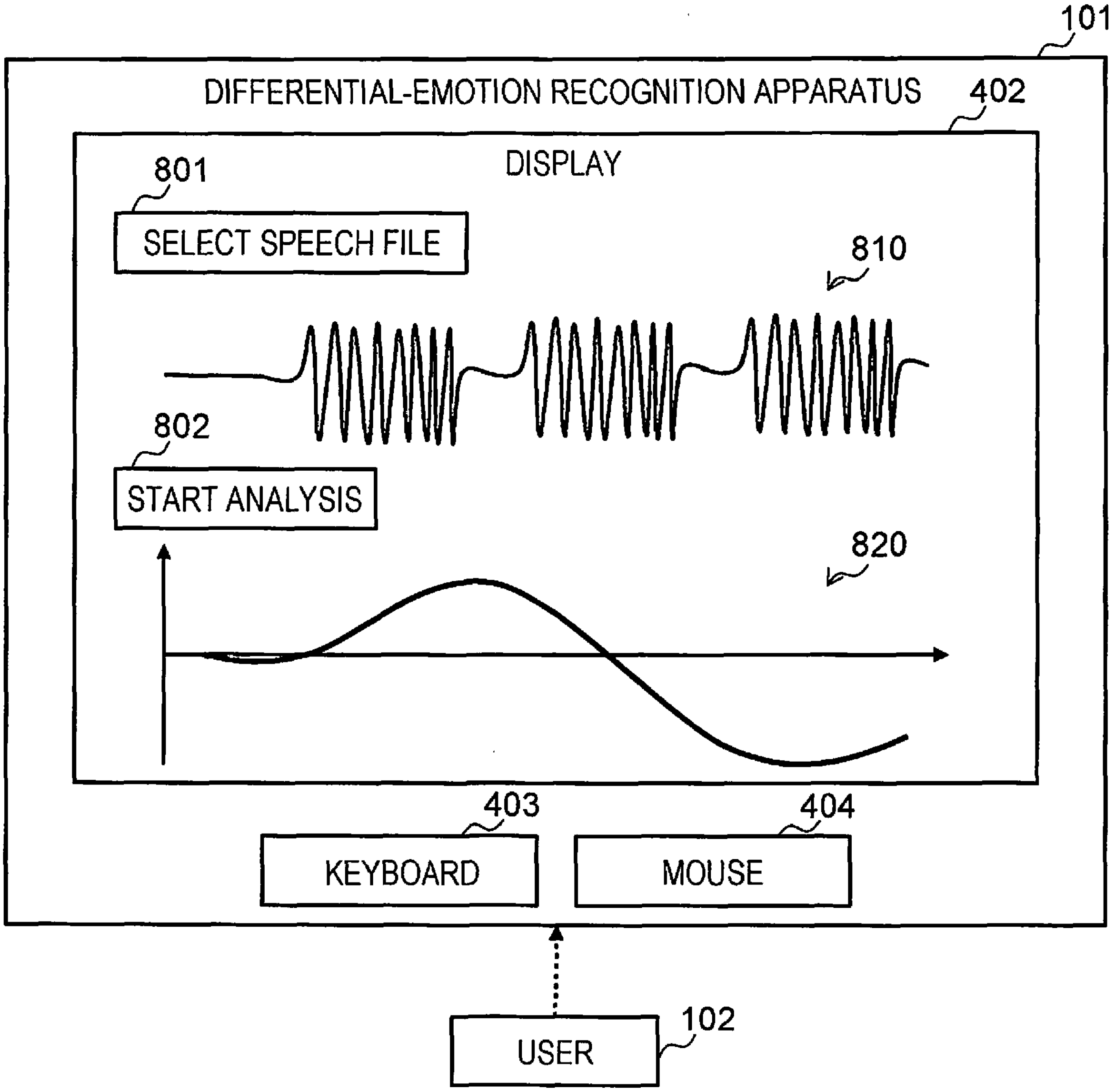
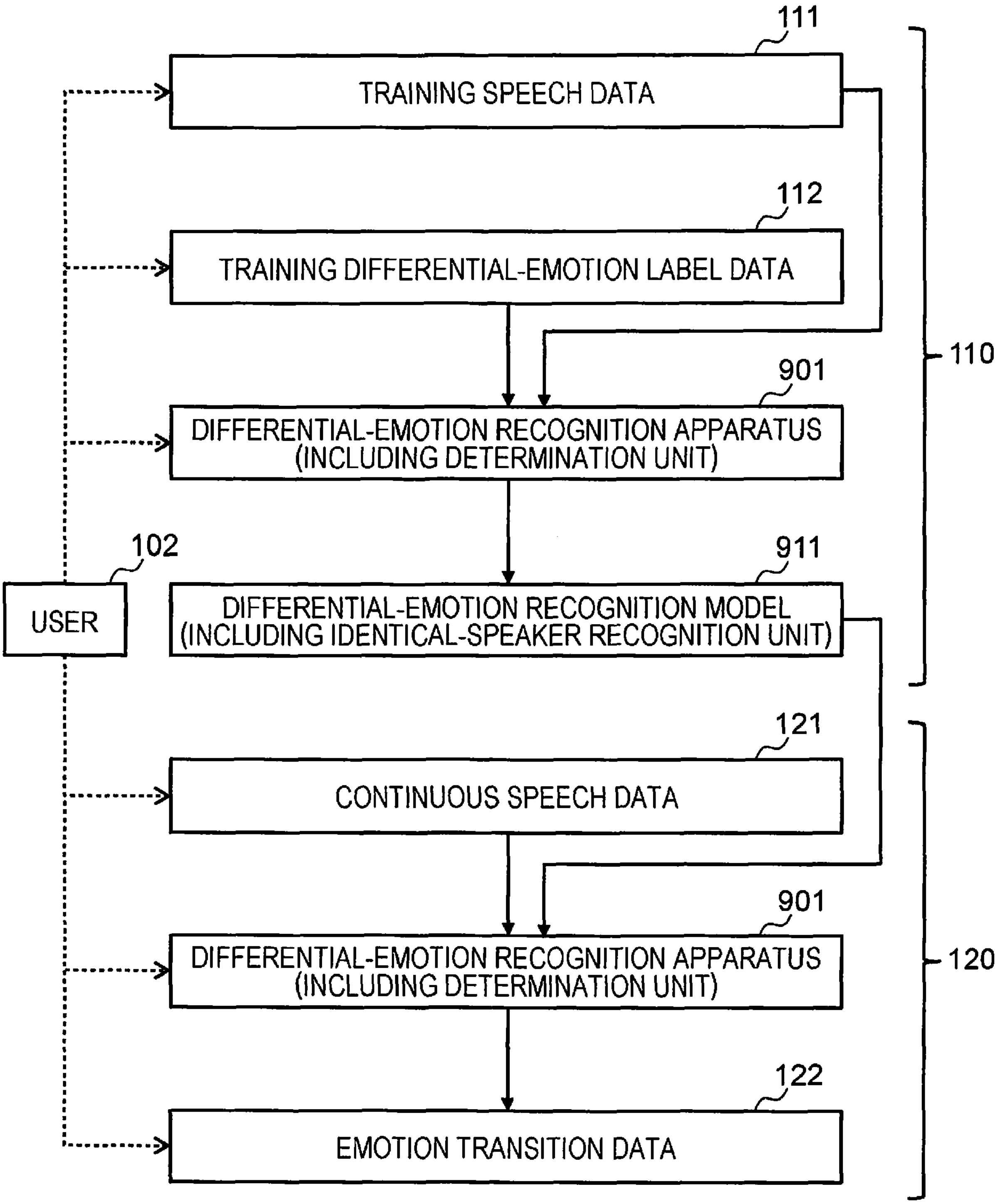

101
DIFFERENTIAL-EMOTION RECOGNITION APPARATUS
402
DISPLAY
801
SELECT SPEECH FILE
810
802
START ANALYSIS
820
403
KEYBOARD
404
MOUSE
USER
102

FIG. 9


111
TRAINING SPEECH DATA
112
TRAINING DIFFERENTIAL-EMOTION LABEL DATA
901
DIFFERENTIAL-EMOTION RECOGNITION APPARATUS (INCLUDING DETERMINATION UNIT)
110
102
USER
911
DIFFERENTIAL-EMOTION RECOGNITION MODEL (INCLUDING IDENTICAL-SPEAKER RECOGNITION UNIT)
121
CONTINUOUS SPEECH DATA
901
DIFFERENTIAL-EMOTION RECOGNITION APPARATUS (INCLUDING DETERMINATION UNIT)
120
122
EMOTION TRANSITION DATA

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 22 20 5114

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
X	US 2020/075040 A1 (PROVOST EMILY MOWER [US] ET AL) 5 March 2020 (2020-03-05) * paragraphs [0006], [0019], [0022], [0041], [0044], [0062] * -----	1-14	INV. G10L25/30 G10L25/63
A	US 9 300 790 B2 (GAINSBORO JAY LORING [US]; WEINSTEIN LEE DAVIS [US] ET AL.) 29 March 2016 (2016-03-29) * column 11, line 67 - column 12, line 26 * -----	2,3,9,10	
			TECHNICAL FIELDS SEARCHED (IPC)
			G10L

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
The Hague	13 February 2023	Taddei, Hervé

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons
....
& : member of the same patent family, corresponding document

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO.

EP 22 20 5114

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report.
The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

13-02-2023

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2020075040 A1	05-03-2020	NONE	
US 9300790 B2	29-03-2016	US 10084920 B1	25-09-2018
		US 2007071206 A1	29-03-2007
		US 2016217807 A1	28-07-2016

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82

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.

Patent documents cited in the description

- JP 2018132623 A **[0004]**

Non-patent literature cited in the description

- **SUZUKI.** Recognition of Emotions Included in Speech. *Journal of the Acoustical Society of Japan,* 2015, vol. 71 (9), 484-489 **[0033]**