(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
01 November 2018 (01.11.2018)

WIPO | PCT

(10) International Publication Number
WO 2018/199997 A1

(51) International Patent Classification:
G06F 15/18 (2006.01) *H04R 1/20* (2006.01)

(21) International Application Number:
PCT/US2017/030213

(22) International Filing Date:
28 April 2017 (28.04.2017)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **HEWLETT-PACKARD DEVELOPMENT COMPANY, L.P.** [US/US]; 11445 Compaq Center Drive W., Houston, Texas 77070 (US).

(72) Inventors: **BHARITKAR, Sunil**; 1501 Page Mill Road, Palo Alto, California 94304 (US). **ATHREYA, Madhu Sudan**; 1501 Page Mill Road, Palo Alto, California 94304 (US).

(74) Agent: **BURROWS, Sarah E.** et al.; HP Inc., 3390 E. Harmony Road, Mail Stop 35, Fort Collins, Colorado 80528 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ,

WO 2018/199997 A1

(54) Title: AUDIO CLASSIFCATION WITH MACHINE LEARNING MODEL USING AUDIO DURATION

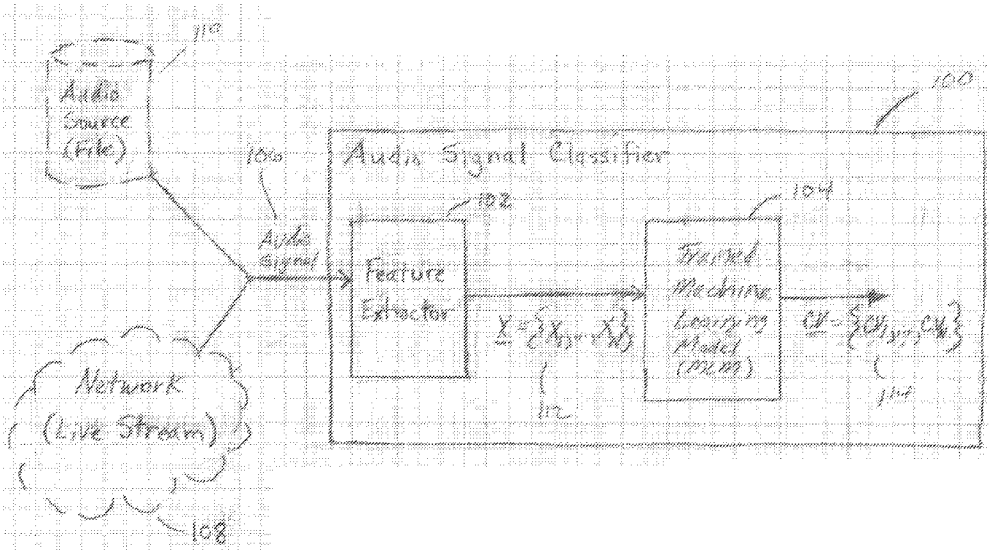

Fig. 1

(57) Abstract: An audio signal classifier including a feature extractor to extract metadata from an audio signal, the metadata defining a plurality of features of the audio signal, the feature extractor to generate a feature vector including selected features of the audio signal, the selected features including a duration of the audio signal, and each selected feature having a feature value. A machine learning model trained to classify the audio signal as one of a plurality of audio signal classes based on the feature vector. The machine learning model to provide a plurality of class values based on the feature values, each class value corresponding to one of the plurality of audio signal classes, the plurality of class values together indicating the class of the audio signal.

[Continued on next page]

TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to the identity of the inventor (Rule 4.17(i))*

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

Published:

— *with international search report (Art. 21(3))*

AUDIO CLASSIFCATION WITH MACHINE LEARNING MODEL USING AUDIO DURATION

Background

[0001] Electronic devices employing loudspeakers (e.g., personal electronic devices, such as cell phones) may include frequency control (e.g., bass, mid-range, and treble frequency control) to adjust audio output from the loudspeakers to improve the quality of experience (QoE) of content involving audio or speech. Audio signals may comprise different classes of audio content, such as music, voice, and movie content, for example, where each class may require different frequency control for optimizing QoE.

Brief Description of the Drawings

[0002] Figure 1 is a block and schematic generally illustrating an audio signal classifier, according to one example.

[0003] Figure 2 is a schematic diagram generally illustrating a machine learning model, according to one example.

[0004] Figure 3 is a block and schematic generally illustrating an audio signal classifier, according to one example.

[0005] Figure 4 is a block and schematic diagram generally illustrating an audio system including an audio signal classifier, according to one example.

[0006] Figure 5 is a table illustrating mean, μ, and standard deviation, σ, for Hollywood cinematic content, according to one example.

[0007] Figure 6 is a histogram illustrating a modeled Gaussian distribution of an example of 500 audio samples of cinematic content, according to one example.

[0008] Figure 7A is a graph illustrating a distribution of YouTube video duration, according to one example.

[0009] Figure 7B is a graph illustrating a distribution of YouTube video duration for music, entertainment, comedy, and sports genres, according to one example.

[0010] Figure 8 is a histogram illustrating a modeled gamma distribution of YouTube sports and comedy content, according to one example.

[0011] Figure 9 is a histogram illustrated a modeled Gaussian distribution of broadcast sports content, according to one example.

[0012] Figure 10 is a graph illustrating mean-squared-error

[0013] Figure 11 is a flow diagram illustrating a method of classifying an audio signal as being of one of a plurality of audio signal classes, according to one example.

[0014] Figure 12 is a flow diagram illustrating a method of classifying an audio signal as being of one of a plurality of audio signal classes, according to one example.

[0015] Figure 13 is a block and schematic diagram generally illustrating a computing system for implementing an audio signal classifier, according to one example.

Detailed Description

[0016] In the following detailed description, reference is made to the accompanying drawings which form a part hereof, and in which is shown by way of illustration specific examples in which the disclosure may be practiced. It is to be understood that other examples may be utilized and structural or logical changes may be made without departing from the scope of the present disclosure. The following detailed description, therefore, is not to be taken in a limiting sense, and the scope of the present disclosure is defined by the appended claims. It is to be understood that features of the various examples

described herein may be combined, in part or whole, with each other, unless specifically noted otherwise.

[0017] Electronic devices (e.g., personal electronic devices, such as cell phones) typically include loudspeakers for playing audio content. Such electronic devices may include a number of audio control presets for adjusting elements of the audio reproduced by the loudspeakers (e.g., bass, mid-range, and treble frequency presets) so as to improve the quality of experience (QoE) of audio content for a user.

[0018] Audio signals may be of a number of different classes of audio content, such as music, voice, and cinema (movie) content, for example, where each audio class may require control of different audio control presets for optimizing QoE for a user. For example, the type of presets may be different for each class of audio signals, with one class requiring three presets (e.g., bass, mid-range, and treble presets) and another class requiring four or more presets (e.g., bass-1, bass-2, mid-range, and treble presets), for instance. Also, classes of audio content using a same set of presets may require different values for each preset.

[0019] Electronic devices often include one or more sets of pre-programmed audio presets for controlling elements of audio output (e.g. bass, mid-range, and treble frequency). For example, some electronic devices enable a user to select one of three pre-programmed sets of presets, one each for music, voice and cinema content. Often, a user is not aware that the pre-programmed sets presets even exist, such that the default preset being used by the device may or may not correspond to the class of audio content being reproduced. Additionally, even if a user is aware of the pre-programmed sets of presets, the user needs to manually select the appropriate set of presets, and manually select an appropriate value for each preset of the set of presets each time different audio content is being reproduced. Such a process in inherently error prone due to a user potentially not being aware of the presets, a user forgetting to apply presets, and a user applying the wrong set of presets and/or wrong preset values to the audio content.

[0020] The present disclosure provides an automated audio signal classifier that can be employed by electronic devices to classify an audio signal as one of a plurality of types or classes of audio content (e.g., voice, music, cinema etc.). The classification of the audio signal is then used to automatically identify and apply proper audio presets to control the audio content being reproduced by the loudspeakers. Such process ensures that optimal audio presets are applied so as to provide an optimal QoE for a user.

[0021] In one example, as will be described in greater detail below, an audio signal classifier, in accordance with the present disclosure, uses features of the audio signal included in metadata of the audio signal or stream to classify the audio signal as one of a plurality of audio signal classifications or types using a trained machine learning model (e.g., a neural network). In one example, among other features, the features include a duration of the audio signal, with the trained machine learning model being trained to classify audio signals based, in part, on the duration of the audio signal.

[0022] Figure 1 is a block and schematic diagram generally illustrated an audio signal classifier 100 including a feature extractor 102 and a trained machine learning model 104, according to one example. In one example, as will be describe in greater detail below, trained machine learning model 104 comprises a neural network. In Figure 1, feature extractor 102 receives a digital audio signal 106, such as in the form of streaming video from a network 108 (e.g., MPEG-2 transport stream, MPEG-4 audio-video file containers), or a file from an audio source 110, such as a database or some type of storage device, where each audio signal 106 can be classified as being one of a plurality of audio signal classes (e.g., voice, music, cinema), and where each audio signal 106 includes metadata defining parameters or features of the audio signal such as, for example, a sample rate of the audio in kHz (e.g., 16, 44.1, 48), duration of the audio signal in seconds, bit-depth in bits/sample (e.g., 16, 20, 24), file size (e.g., kilobytes), bitrate (e.g., bits/second), presence/absence of video content (video-bit 0 or 1), audio channel count (e.g., 1, 2, 6, 8), and presence object-based audio or channel-based audio (e.g., {0, 1}).

[0023] In one example, feature extractor 104 generates a feature vector, $\underline{X}$, for audio signal 106, as indicated at 112, where feature vector $\underline{X}$ includes a plurality of audio features, indicated as audio features X_1 to X_N ($\underline{X}=\{X_1,\ldots,X_N\}$) selected from the metadata of audio signal 106. In one example, the selected audio features at least include the duration of the audio signals in seconds. In one instance, where metadata for audio signal 106 does not explicitly include a duration of the audio signal, feature extractor 104 generates and includes duration feature in feature vector $\underline{X}$ based on the file size and bitrate features (e.g., file size (kilobytes)/bitrate (bits/second)). The file-size information can be obtained from the file-data and duration can be computed using file-size/bitrate of the content, or can be determined by computing the difference in the beginning and end time-stamps of the file or stream. In one example, feature vector $\underline{X}$ includes six features (i.e., $\underline{X}=\{X_1, \ldots,X_6\}$), the six selected features being: duration of the audio signal in seconds, sample rate of the audio in kHz (e.g., 16, 44.1, 48), bit-depth in bits/sample (e.g., 16, 20, 24), presence or absence of video content (video-bit 0 or 1), audio channel count (e.g., 1, 2, 6, 8), and presence object-based audio or channel-based audio (e.g., {0, 1}).

[0024] According to one example, trained machine learning model 104 is trained to classify audio signal 106 as one of a plurality of predefined audio classes based on feature vector $\underline{X}$. In one example, trained machine learning model is trained to classify audio signal as being one of three classes of audio signals (i.e., voice, music, and cinema). In one example, as described in greater detail below, trained machine learning model 104 is trained using training or sample vectors $\underline{X}$ constructed to represent a statistical distribution of actual audio content.

[0025] In one example, trained machine learning model 104 receives feature vector $\underline{X}$, and, based on the values of features X_1 to X_N, provides a plurality of class output values $\underline{CV}$, as indicated at 114, and illustrated as CV_1 to CV_X, (e.g., $\underline{CV} = \{CV_1,\ldots,CV_X\}$), where each class output value CV_1 to CV_X corresponds to different one of the plurality of audio classes. According to one example, classes are substantially non-overlapping due to the choice of feature vector, but may be partially overlapping depending on whether new features are added

or deleted from the feature vector set, and the output values CV_1 to CV_X are substantially separable based on a threshold criteria. In one example, as described above, trained machine learning model 104 provides three class output values, CV_1 to CV_3, respectively corresponding to voice, music, and cinema audio classes.

[0026] According to one example, the plurality of class values CV_1 to CV_X together are indicative of the class of the audio signal. In one example, as will be described in greater detail, the class of the audio signal, as automatically identified by audio signal classifier 100, is used to automatically identify and apply proper audio presets to control the audio content being reproduced by the loudspeakers of an audio system so as to optimize QoE for a user.

[0027] Figure 2 is a schematic diagram generally illustrating an example of machine learning model 104, where machine learning model 104 is implemented as a neural network. In one example, machine learning model 104 includes an input layer 120 including a plurality of input neurons 122, one input neuron 122 corresponding to and receiving a different one of the audio features X_1 to X_N of feature vector $\underline{X}$, and an output layer 124 including a plurality of output neurons 124, one output neuron 124 corresponding to and providing a different one of the output class values C_1 to C_N. In one example, machine learning model 104 includes a plurality of hidden neural layers 130, such as hidden layer 132 including a number of neurons 134 and hidden layer 136 including a number of neurons 138, with hidden layers 130 interconnected by a plurality of synapses, such as synapse 140, between input and output layers 130.

[0028] In one example, machine learning model 104 includes an input layer 120 having six input neurons 122, one input neuron 122 corresponding to a different one of the six audio features of feature vector $\underline{X}$ as described above (i.e., duration, sample rate, bit-depth, presence or absence of video content, audio channel count, and presence of object-based audio or channel-based audio), an output layer 124 including three output neurons 126, one output neuron 126 corresponding to a different one of the three audio classes described above

(i.e., voice, music, and cinema classes), and two hidden layers, such as hidden layers 132 and 134, where each hidden layer includes 10 hidden neurons.

[0029] In one example, machine learning model 104 includes an input layer 120 having six input neurons 122, one input neuron 122 corresponding to a different one of the six audio features of feature vector $\underline{X}$ as described above (i.e., duration, sample rate, bit-depth, presence or absence of video content, audio channel count, and presence of object-based audio or channel-based audio), an output layer 124 including three output neurons 126, one output neuron 126 corresponding to a different one of the three audio classes described above (i.e., voice, music, and cinema classes), and two hidden layers, such as hidden layers 132 and 134, where each hidden layer includes 10 hidden neurons.

[0030] In one example, trained machine learning model 104 may employ any of a number of processing techniques such as, for example, a Bayesian Classifier, MLP with gradient descent based learning, etc. Based on the input feature vector and the corresponding associated labeled class value the output neuron produces a value. The MLP is trained on the sum-squares errors (difference between the neuron outputs and the desired output). For example if the feature-vector corresponds to movie feature set, then the output class values will be CV1=1, CV2=-1, CV3=-1 (where CV1 corresponds to the neuron or output for movie). The error is computed from the output of the three output neurons and the weights are adapted using the gradient descent algorithm. The next feature vector is delivered to the network and the error computed based on the output of the neurons and the desired class output values and the weights adapted to minimize the error. The process is repeated for all feature vectors and the feature vectors are repeatedly presented multiple times until the error is minimized (example of the error plot is shown in Fig.10).

[0031] Figure 3 is a block and schematic diagram generally illustrating an example implementation of audio signal classifier 100. According to the example of Figure 3, in addition to trained machine learning model 104, audio signal classifier 100 includes a trained deep learning model 154 which is employed or “switched in” when audio signal 106 is determined by a feature evaluator 140 to have confounding or invalid metadata (e.g., metadata is

missing, there is contradictory metadata, the metadata has abnormal values, etc.) or when output class values CV generated by trained machine learning model 104 are determined by a reliability evaluator 142 to be unreliable ((i.e., the output class values do not provide a clear indication as to the class of the audio signal). With the inclusion of trained deep learning model 154, the example audio signal classifier 100 of Figure 3 may be referred as a dual-model machine learning audio signal classifier.

[0032] In contrast to trained machine learning model 104, which classifies audio signal 106 based on metadata from audio signal 106, trained deep learning model 154 classifies audio signal 106 based on decoded audio frames from audio signal 106 (e.g. time-domain frames and/or time frequency data computed using short-time Fourier transforms (STFT) over frames (e.g., 20 ms of audio data)). In one example, trained deep learning model 154 comprises a neural network employing multi-stage classifiers. In one example, trained deep learning model 154 is trained on frames of labeled audio data such as, for example, explosions, applause, Foley (i.e., reproduced sound effects), music, etc. Based on the decoded audio frames, trained deep learning model 154 outputs a plurality of output class values, CV', with each class value corresponding to a different class of the plurality of audio classes (e.g., voice, music, and cinema), and the plurality of output class values together indicating the class of audio signal 106.

[0033] Examples of the operation of audio signal classifier 100 of Figure 3 are described below. Initially, feature extractor 102 receives audio signal 106 and extracts metadata therefrom. According to one example, feature evaluator 140 evaluates the integrity or validity of the metadata (e.g. whether there is metadata missing, whether there is contradictory metadata, whether the metadata has atypical values, etc.). In one case, feature evaluator 40 generates a robustness value, D, having a value of either “0” or “1” (D:{0, 1}) based on the extracted metadata. In one example, D has a value of “0” (D=0) when the metadata is valid, and a value of “1” (D=1) when the reliability of the metadata is confounding (e.g., when there is missing metadata, corrupted metadata, contradictory metadata, or atypical metadata).

[0034] According to the example of Figure 3, feature extractor 102 provides decoded audio frames 144 to an audio input controller 146. In one example, audio input controller 146 either passes decoded audio frames 144 to trained deep learning model 154 for processing or blocks trained deep learning model 154 from receiving decoded audio frames 144, depending on robustness value, D, generated by feature evaluator 140 and on a reliability value, β, generated by reliability evaluator 142 (which will be described in greater detail below. In one example, audio input controller 146 passes decoded audio frames 144 to trained deep learning model 154 by applying a gain with a value of "1" to decoded audio frames 144, or blocks trained deep learning model 154 from receiving decoded audio frames 144 by applying a gain having a value of "0" to decoded audio frames 144.

[0035] Continuing with the operation of audio signal classifier 100, when robustness value D=0, feature extractor 102 provides feature vector X to trained machine learning model 104. In response, trained machine learning module 104 provides the plurality of output class values CV (e.g., one class value for each class of a plurality of audio classes) to reliability evaluator 142 and to a corresponding MLM (machine learning model) decision model 148. Additionally, it is noted that audio input controller 146 does not pass decoded audio frames 144 to trained deep learning model 154 in response to robustness value D being "0".

[0036] In one example, upon receiving output class values CV, reliability evaluator 142 generates a reliability index, α, which is indicative of the reliability of output class values CV (i.e., how reliable or accurate will the resulting classification be based on such output class values). In one case, the reliability index, α, is based on an amount of separation between the class values CV. In one example, reliability index α is the root mean square error between each of the output class values CV. In one example, if the reliability index α is greater than or equal to a threshold value, T, the output class values CV are deemed to be reliable, and reliability evaluator 142 provides a reliability value, β, having a value of "1" (β=1). Conversely, if the reliability index α is less than the threshold

value, T, reliability evaluator 142 provides a reliability value, β, having a value of "1" (β=0), indicating that class values CV are deemed to be unreliable.

[0037] In a scenario where the β=1 (meaning that output class values CV are reliable), audio input controller 146 does not pass decoded audio frames 144 to trained deep learning model 154. Additionally, with β=1, MLM decision model 148 determines the class of audio signal 106 based on the plurality of output class values CV. In one case, MLM decision model 148 classifies audio signal 106 as belonging to the audio class corresponding to the class value of the plurality of class values CV having the highest value. For example, in a case where the plurality of audio classes are {movie, music, voice} and the corresponding CV values are CV={-1, +1, -1}, MLM decision model 148 will classify audio signal 106 as "music", since music has the highest corresponding class value (i.e. "+1"). According to this scenario, where β=1, MLM decision model 148 passes the determined audio class (e.g., "movie") to global decision model 150 which, in this case, acts as a "pass-thru" and provides the identified audio class received from MLM decision model 148 as output audio class 152. In one example, output audio class 152 is used to select audio presets to adjust an audio output of loudspeakers (e.g., see Figure 4 below).

[0038] In a case where the β=0 (meaning that output class values CV are not reliable), rather than determining the audio class of audio signal 106 and providing an identified audio class to global decision model 150, MLM decision model 148 instead passes the plurality of output class values CV to global decision model 150. An example of this is a case when the CV values are distributed as {0.2, -0.1, -0.7} where the separation between movie and music class values are not significant (significance being determined based on pairwise error computation between class values). Additionally, with β=0, audio input controller 146 passes decoded audio frames 144 to trained deep learning model 154. In response, trained deep learning model 154 generates and provides a plurality of output class values CV' to a DLM decision model 156 corresponding to trained deep learning model 154, where each output class value corresponds to a different class of the plurality of audio classes. With β=0, rather than determining the audio class of audio signal 106 and providing

an identified audio class to global decision model 150, DLM decision model 156 passes the plurality of output class values CV′ to global decision model 150.

[0039] In response to receiving the plurality of output class values CV and the plurality of output class values CV′, global decision model 150 does not act as a pass-thru, but instead determines an audio class for audio signal 106 based on the two sets of output class values. Global decision model 150 may employ any number of techniques for determining an audio class for audio signal 106. In one case, global decision model 50 simply classifies audio signal 106 as belonging to the audio class corresponding the class values having the largest sum. For example, in a case where the plurality of audio classes are {movie, music, voice} and the corresponding CV values are CV={0.5, 0.4, 0.1} and CV′ values are CV′={0.6, 0.1, 0.3}, global decision model 150 will designate audio signal 106 as a "movie" since the sum of the corresponding class values has the highest value (i.e., {movie, music, voice}= {1.1, 0.5, 0.4}).

[0040] In another example, global decision model 150 may employ a linear weighted average. For example, global decision model 150 may apply a "weight1" to the plurality of class values CV, and a "weight2" to the plurality of class values CV, such that movie=((0.5*weight1+0.6*weight2)/(weight1 + weight2)); music=((0.4*weight1+0.1*weight2)/(weight1 + weight2)); and voice=((0.1*weight1+0.3*weight2)/(weight1 + weight2)). If weight1=0.5 and weight2=1, then {movie, music, voice} = {0.57, 0.2, 0.23}, such that global decision model 150 will designate audio signal 106 as a "movie".

[0041] Returning to feature evaluator 140, in a scenario where robustness value, D, has a value of "1" (D=1), meaning that metadata has been deemed to be unreliable, audio input controller 146 will pass decoded audio frames 144 to trained deep learning model 154. However, since trained machine learning model 104 is not trained on unreliable metadata (i.e., unreliable feature values), feature extractor 102 does not provide feature vector X to trained deep learning model 104.

[0042] In such scenario, upon receiving decoded audio frames 144 via audio input controller 146, trained deep learning model 154 generates and provides the plurality of output class values CV′ to DLM decision model 156. With D=1,

DLM decision model 156 determines the class of audio signal 106 based on the plurality of output class values $\underline{CV'}$. In one case, DLM decision model 156 classifies audio signal 106 as belonging to the audio class corresponding to the class value of the plurality of class values $\underline{CV'}$ having the highest value. For example, in a case where the plurality of audio classes are {movie, music, voice} and the corresponding $\underline{CV'}$ values are $\underline{CV'}$ = {-1, +1, -1}, DLM decision model 156 will classify audio signal 106 as "music", since music has the highest corresponding class value (i.e. "+1"). DLM decision model 156 passes the determined audio class (e.g., "movie") to global decision model 150 which, in this case (D=1), acts as a "pass-thru" and provides the identified audio class received from DLM decision model 156 as output audio class 152.

[0043] In view of the above, when D=0 and β=0, audio classifier 100 of Figure 3 employs only trained machine learning model 104 to determine an audio class of audio signal 106. Conversely, when D=1, audio classifier 100 of Figure 3 employs only deep trained learning model 154 to determine an audio class of audio signal 106. Finally, when D=0 and β=1, audio classifier 100 of Figure 3 employs both trained machine learning model 104 and trained deep learning model 154 to determine an audio class of audio signal 106. In one example, MLM decision model 148, DLM decision model 156, and global decision block 150 together form and output decision model 158.

[0044] Figure 4 is a block and schematic diagram generally illustrating an audio system 180 including a loudspeaker system 182 and an audio signal classifier 100, such as described by Figures 1-3, according to one example. Loudspeaker system 182 includes a plurality of sets of audio presets, each corresponding to a different audio class, and one or more loudspeakers 186 for reproducing audio signal 106. In one example, such as described above with reference to Figures 1-3, audio classifier 100 classifies audio signal 106 as belong to one class of a plurality of audio signal classes (e.g., voice, music, cinema) and provides indication of the identified audio class 190 of audio signal 106 to loudspeaker system 182. In one example, loudspeaker system 182 selects the set of audio presets corresponding to the identified audio class to adjust the audio output of loudspeakers 186.

[0045] To employ the duration feature of an audio signal as a classifying feature, since content length can vary significantly, the duration of the audio signal is modeled. According to one example, statistical distributions are used to model duration for audio content (e.g., voice, music and cinema).

[0046] Figure 5 is a table showing publicly available information regarding mean, μ, and standard deviation, σ, for Hollywood cinematic content. Given the mean and standard deviation represent second-order statistics of normal distributions, according to one example, 500 samples of duration data for training samples were generated using the mean and standard deviation of Figure 5 and Equation I as follows:

Equation I: $$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{(-\frac{(x-\mu)^2}{2\sigma^2})}.$$

[0047] Figure 6 is a histogram showing a modeled Gaussian distribution of an example of 500 audio samples used for cinematic (movie) content using the distribution of Equation I above.

[0048] Similarly, content distributions of durations for publically available YouTube content are shown in Figure 7A. In one example, for YouTube music distribution, a Gaussian distribution with appropriate mean, μ, and standard deviation, σ, using Equation I was applies to generate 500 samples, whereas for sports and comedy distribution (labeled as "voice"), a gamma distribution according to Equation II below was used to generate 500 samples for these two classes using the gamma function Γ(α),

Equation II: $$f(x|a,b) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{(-x/b)}.$$

[0049] Distributions for durations of YouTube content for music, entertainment, comedy, and sports, as described above, are illustrated by Figure 7B.

[0050] A modeled Gaussian distribution of YouTube sports broadcast content is illustrated by Figure 8, where again, appropriate mean, μ, and standard deviation, σ, were employed using Equation II to generate 500 training samples for machine learning model 104.

[0051] In one example, samples generated from distribution modeling of the duration were permuted with other features of the feature vector (e.g., sample rate, bit depth, number of channels, video presence) to create 500 training feature vectors in a meaningful way based on how typical audio content is encoded an exists. In one example, the training feature vectors were randomized before applying them to machine learning model 104, with the training be done to minimize the sum-squares error (e.g., the difference between actual output of the output neuron for each class and a value labeled for a target class, either a -1 or a +1 for a hyperbolic tangent transfer function of the output neuron) over all outputs and training samples using the Levenberg-Marquart algorithm for updating the synapse weights of the machine learning model. It is noted that a sigmoid with output values $\in [0,1]$ does not change classification accuracy.

[0052] Figure 10 is a graph illustrating exemplary results of training a machine learning model 104 of audio signal classifier 100, as illustrated by Figure 2, where 6 input neurons 122 were employed, two hidden layers 132 and 134 were used, with each hidden layer using 10 neurons, and an output layer 124 having three output neurons 126 was used. Each of the 6 input neurons received a different one of six feature values (X_1, …,X_6) of feature vector X (i.e., duration of the audio signal in seconds, sample rate of the audio in kHz (e.g., 16, 44.1, 48), bit-depth in bits/sample (e.g., 16, 20, 24), presence or absence of video content (video-bit 0 or 1), audio channel count (e.g., 1, 2, 6, 8), and presence object-based audio or channel-based audio (e.g., {0, 1})), and each of the three output neurons 126 provided a class value for a different class of three possible audio signal classes (i.e., voice, music, and cinema). In Figure 10, it is noted that the 3 curves substantially merge and overlay one another below 300 epochs.

[0053] Exemplary classification results for several actual cinematic, sports (voice), and music videos using the above-described trained machine learning model 104 are described below. For the movie “Edge of Tomorrow” having a duration of 6,780 seconds, a sample rate of 48 kHz, an audio channel count of 8, a bit-depth of 24, a video bit = 1, and an object bit =0, trained machine

learning model provided class output values of +1.0 for the movie class, -0.99 for the music class, and -1.0 for the voice class. Based on maxima, the trained machine learning model 104 correctly identified the audio signal as being of the movie class.

[0054] As another example, for the movie "Batman (The Dark Knight Rises)" having a duration of 9,900 seconds, a sample rate of 48 kHz, an audio channel count of 6, a bit-depth of 24, a video bit = 1, and an object bit =0, trained machine learning model 104 provided class output values of +1.0 for the movie class, -1.0 for the music class, and -1.0 for the voice class. Based on maxima, the trained machine learning model 104 correctly identified the audio signal as being of the movie class.

[0055] As another example, for a "YouTube music video for Maroon-5" having a duration of 61 seconds, a sample rate of 44.1 kHz, an audio channel count of 2, a bit-depth of 16, a video bit = 1, and an object bit =0, trained machine learning model 104 provided class output values of -1.0 for the movie class, +1.0 for the music class, and -1.0 for the voice class. Based on maxima, the trained machine learning model 104 correctly identified the audio signal as being of the music class.

[0056] As another example, for a "YouTube sports video of a Georgia vs. North Carolina football game" having a duration of 9440 seconds, a sample rate of 44.1 kHz, an audio channel count of 2, a bit-depth of 16, a video bit = 1, and an object bit =0, the trained machine learning model 104 provided class output values of -1.0 for the movie class, -1.0 for the music class, and +1.0 for the voice class. Based on maxima, the trained machine learning model 104 correctly identified the audio signal as being of the voice class.

[0057] Figure 11 is a flow diagram generally illustrating a method 200 of classifying an audio signal as being of one of a plurality of audio signal classes, according to one example. At 202, metadata is extracted from an audio signal, the metadata defining a plurality of features of the audio signal, such as feature extractor 102 extracting metadata from audio signal 106 as illustrated and described by Figures 1 and 3, for example.

[0058] At 204, a feature vector is generated which includes selected features of the audio signal, the selected features including a duration of the audio signal, each selected feature having a feature value, such feature extractor 102 generating a feature vector X from metadata of audio signal 106 as illustrated and described by Figure 1 and 3, for example.

[0059] At 206, method 200 includes generating a plurality of class values based on the feature values of the feature vector using a trained machine learning model, such as trained machine learning model 104 generating output class values CV, as described with respect to Figures 1 and 3, where each class value corresponds to different one of a plurality of audio signal classes (e.g., voice, music, cinema), and where the plurality of class values together indicate the class of the audio signal, the class of the audio signal to select audio presets to adjust audio output of loudspeakers (see Figure 4, e.g.).

[0060] Figure 12 is a flow diagram generally illustrating a method 220 of classifying an audio signal as being of one of a plurality of audio signal classes, according to one example. At 222, metadata is extracted from an audio signal, the metadata defining a plurality of features of the audio signal, such as feature extractor 102 extracting metadata from audio signal 106, as described above with respect to Figures 1 and 3, for example.

[0061] At 224, it is queried whether the extracted metadata is reliable (e.g. the metadata is not corrupt, metadata is not missing, metadata does not have atypical values), such as feature evaluator 140 determining a robustness value, D, as illustrated above with respect to Figure 3, for example. According to one example, if the answer to the query at 224 is “yes”, method 220 proceeds to 226.

[0062] At 226, a feature vector is generated from the metadata, the feature vector including selected features of the audio signal, including a duration of the audio signal, each selected feature having a feature value, such as feature extractor 102 generating feature vector X from metadata extracted from audio signal 106. In one example, in addition to a duration of the audio signal, the feature vector includes a plurality of additional features, such as a sample rate, a bit-depth, a presence or absence of video data, an audio channel count, and

presence or absence of object-based audio or channel-based audio, for example.

[0063] At 228, method 220 includes employing a trained machine learning model to generate from the feature vector a plurality of output class values based on the feature values, with each output class value corresponding to one class of the plurality of audio signal classes (e.g., voice, music, cinema), such as trained machine learning model 104 of Figures 1 and 3 generating a first plurality of output class values, CV. In one example, the trained machine learning model comprises a neural network, such as illustrated by Figure 2.

[0064] At 230, it is queried whether the first plurality of output class values generated by the trained machine learning model is reliable, such as reliability evaluator 142 evaluating whether the plurality of output values CV generated by trained machine learning model 104 are valid via generation of validity value, β, as illustrated and described with respect to Figure 3. In one example, if the answer to the query at 230 is "yes", meaning the class values are reliability (e.g. β = 1), method 220 proceeds to 232.

[0065] At 232, an audio class is determined for the audio signal based on the values of the first plurality of class values generated by the trained machine learning model at 230, such as MLM decision model 148 determining an audio class to which input signal 106 belongs based on the plurality of output class values CV generated by trained machine learning model 104, as described above with respect to Figure 3, when robustness value D=0 and reliability value β=1.

[0066] Returning to 230, if the answer to the query at 230 is "no", meaning that the output class values generated by the trained machine learning model are not reliable, method 220 proceeds to 234. At 234, a trained deep learning model generates a second plurality of output class values based on audio frames extracted from the audio signal, such as trained deep learning model 154 generating a set of output class values CV′ based on audio frames 144, as illustrated and described by Figure 3.

[0067] At 236, a class of the audio signal is determined from the first set of output class values generated by the trained machine learning model at 228

and on the second set of output class values generated by the trained deep learning model at 236, such as output class values CV generated by trained machine learning model 104 and output class values CV′ generated by trained deep learning model 154 as illustrated by Figure 3, when robustness value D=0 and β=0.

[0068] Returning to 224, if the query as to whether the metadata is reliable is "no", method 220 proceeds to 238. At 238, a trained deep learning model generates a second plurality of output class values based on audio frames extracted from the audio signal, such as trained deep learning model 154 generating a set of output class values CV′ based on audio frames 144, as illustrated and described by Figure 3.

[0069] At 240, an audio class is determined for the audio signal based on the values of the second plurality of class values generated by the trained deep learning machine learning model, such as DLM decision model 156 of Figure 3 determining an audio class to which input signal 106 belongs based on the plurality of output class values CV′ generated by trained machine learning model 104, as described above with respect to Figure 3, when robustness value D=1.

[0070] In one example, audio signal classifier 100, including feature extractor 102 and trained machine learning model 104, may be implemented by a computing system. In such examples, audio signal classifier 100, including each of the feature extractor 102 and trained machine learning model 104, of the computing system may include any combination of hardware and programming to implement the functionalities of audio signal classifier 100, including global feature extractor 102 and trained machine learning model 104, as described herein in relation to any of FIGS. 1-12. For example, programming for audio signal classifier 100, including feature extractor 102 and trained machine learning model 104, may be implemented as processor executable instructions stored on at least one non-transitory machine-readable storage medium and hardware may include at least one processing resource to execute the instructions. According to such examples, the at least one non-transitory machine-readable storage medium stores instructions that, when executed by

the at least one processing resource, implement audio signal classifier 100, including feature extractor 102 and trained machine learning model 104.

[0071] Figure 13 is a block and schematic diagram generally illustrating a computing system 300 for implementing audio signal classifier 100 according to one example. In the illustrated example, computing system or computing device 300 includes processing units 302 and system memory 304, where system memory 304 may be volatile (e.g. RAM), non-volatile (e.g. ROM, flash memory, etc.), or some combination thereof. Computing device 300 may also have additional features/functionality and additional or different hardware. For example, computing device 300 may include input devices 310 (e.g. keyboard, mouse, etc.), output devices 312 (e.g. display), and communication connections 314 that allow computing device 300 to communicate with other computers/applications 316, wherein the various elements of computing device 300 are communicatively coupled together via communication links 318.

[0072] In one example, computing device 300 may include additional storage (removable and/or non-removable) including, but not limited to, magnetic or optical disks or tape. Such additional storage is illustrated in Figure 13 as removable storage 306 and non-removable storage 308. Computer storage media includes volatile and nonvolatile, removable and non-removable media implemented in any suitable method or technology for non-transitory storage of information such as computer readable instructions, data structures, program modules, or other data, and does not include transitory storage media. Computer storage media includes RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical storage, magnetic cassettes, magnetic tape, and magnetic disc storage or other magnetic storage devices, for example.

[0073] System memory 304, removable storage 306, and non-removable storage 308 represent examples of computer storage media, including non-transitory computer readable storage media, storing computer executable instructions that when executed by one or more processors units of processing units 302 causes the one or more processors to perform the functionality of a system, such as audio signal classifier 100. For example, as illustrated by

Figure 13, system memory 304 stores computer executable instructions 400 for audio signal classifier 100, including feature extractor instructions 402 and trained machine learning model instructions 404, that when executed by one or more processing units of processing units 302 implement the functionalities of audio signal classifier 100, including feature extractor 102 and trained machine learning model 104, as described herein. In one example, one or more of the at least one machine-readable medium storing instructions for audio signal classifier 100, including feature extractor 102 and trained machine learning module 102, may be separate from but accessible to computing device 300. In other examples, hardware and programming may be divided among multiple computing devices.

[0074] In some examples, the computer executable instructions can be part of an installation package that, when installed, can be executed by at least one processing unit to implement the functionality of audio signal classifier 100. In such examples, the machine-readable storage medium may be a portable medium, such as a CD, DVD, or flash drive, for example, or a memory maintained by a server from which the installation package can be downloaded and installed. In other examples, the computer executable instructions may be part of an application, applications, or component already installed on computing device 300, including the processing resource. In such examples, the machine readable storage medium may include memory such as a hard drive, solid state drive, or the like. In other examples, the functionality of audios signal classifier 100, including feature extractor 102 and trained machine learning model 104 may be implemented in the form of electronic circuitry.

[0075] Although specific examples have been illustrated and described herein, a variety of alternate and/or equivalent implementations may be substituted for the specific examples shown and described without departing from the scope of the present disclosure. This application is intended to cover any adaptations or variations of the specific examples discussed herein. Therefore, it is intended that this disclosure be limited only by the claims and the equivalents thereof.

CLAIMS

1. An audio signal classifier comprising:

a feature extractor to extract metadata from an audio signal, the metadata defining a plurality of features of the audio signal, the feature extractor to generate a feature vector including selected features of the audio signal, the selected features including a duration of the audio signal, and each selected feature having a feature value; and

a machine learning model trained to classify the audio signal as one of a plurality of audio signal classes based on the feature vector, the machine learning model to generate a plurality of class values based on the feature values, each class value corresponding to one of the plurality of audio signal classes, the plurality of class values together indicating the class of the audio signal, the class of the audio signal to select audio presets to adjust audio output of loudspeakers.

2. The audio signal classifier of claim 1, further including:

a deep learning model trained with a plurality of modeled audio frames each representing a different sound of a plurality of sounds, the deep learning model to generate a plurality of class values based on audio frames of the audio signal, the plurality of class values together indicating the class of the audio signal, the class of the audio signal to select audio presets to adjust audio output of loudspeakers.

3. The audio signal classifier of claim 2, the feature extractor to generate a robustness value to indicate whether the extracted metadata is valid or invalid, the audio signal classifier further including:

a reliability evaluator to generate a reliability value to indicate whether the plurality of class values generated by the machine learning model is reliable or unreliable.

4. The audio signal classifier of claim 3, including an output decision model to determine a class of the audio signal from:

only the plurality of class values generated by the machine learning model when the robustness value indicates that the extracted metadata is valid and when the reliability value indicates that the plurality of class values generated by the machine learning model is reliable;

only the plurality of class values generated by the deep learning model when the robustness value indicates that the extracted metadata is invalid; and

the plurality of class values generated by the machine learning model and the plurality of class values generated by the deep learning model when the robustness value indicates that the extracted metadata is valid and the reliability value indicates that the plurality of class values generated by the machine learning model is unreliable.

5. The audio signal classifier of claim 1, the feature vector, in addition to the duration of audio signal, including the selected features of a sample rate, a bit-depth, a presence or absence of video data, an audio channel count, and a presence or absence of object-based or channel-based audio.

6. The audio signal classifier of claim 1, the plurality of audio signal classes comprising a voice class, a music class, and a movie class.

7. The audio signal classifier of claim 1, the machine learning model comprising a neural network including:

a plurality of input neurons, each input neuron corresponding to a different one of the selected features of the feature vector; and

a plurality of output neurons, each output neuron providing a class value corresponding to a different one of the plurality of audio classes.

8. A non-transitory computer-readable storage medium comprising computer-executable instructions, executable by at least one processor to:

implement a feature extractor to:

extract metadata from an audio signal, the metadata defining a plurality of features of the audio signal; and

generate a feature vector including selected features of the audio signal, the selected features including a duration of the audio signal, each selected feature having a feature value; and

implement a trained machine learning model to:

generate a plurality of class values based on the feature values of the feature vector, each class value corresponding to a different class of a plurality of audio signal classes, the plurality of class values together indicating the class of the audio signal, the class of the audio signal to select audio presets to adjust audio output of loudspeakers.

9. The non-transitory computer-readable storage medium of claim 8, further including computer-executable instructions, executable by the at least one processor to:

implement a deep learning model to:

generate a plurality of class values based on audio data from audio frames of the audio signal, each class value corresponding to a different class of the plurality of audio signal classes, the plurality of class values together indicating the class of the audio signal

10. The non-transitory computer-readable storage medium of claim 9, further including computer-executable instructions, executable by the at least one processor to:

implement the feature extractor to:

generate a robustness value to indicate whether the extracted metadata is valid or invalid; and

implement a reliability evaluator to generate a reliability value to indicate with the plurality of class values generated by the machine learning model is reliable or unreliable.

11. The non-transitory computer-readable storage medium of claim 9, further including computer-executable instructions, executable by the at least one processor to:

implement and output decision model to determine a class of the audio signal from:

only the plurality of class values generated by the machine learning model when the robustness value indicates that the extracted metadata is valid and when the reliability value indicates that the plurality of class values generated by the machine learning model is reliable;

only the plurality of class values generated by the deep learning model when the robustness value indicates that the extracted metadata is invalid; and

the plurality of class values generated by the machine learning model and the plurality of class values generated by the deep learning model when the robustness value indicates that the extracted metadata is valid and the reliability value indicates that the plurality of class values generated by the machine learning model is unreliable.

12. A method of classifying audio signals comprising:

extracting metadata from an audio signal, the metadata defining a plurality of features of the audio signal;

generating a feature vector including selected features of the audio signal, the selected features including a duration of the audio signal, each selected feature having a feature value; and

generating a first plurality of class values based on the feature values of the feature vector with a trained machine learning model, each class value corresponding to different class of a plurality of audio signal classes, the first plurality of class values together indicating the class of the audio signal, the class of the audio signal to select audio presets to adjust audio output of loudspeakers.

13. The method of claim 12, including:

generating a second plurality of class values based on audio frames of the audio signal with a deep learning model, each class value corresponding to different class of the plurality of audio signal classes, the first plurality of class values together indicating the class of the audio signal.

14. The method of claim 13, including:

generating a robustness value indicating whether the extracted metadata is valid or invalid; and

generating a reliability value indicating wither the first plurality of class values is reliable or unreliable.

15. The method of claim 14, including determining a class of the audio signal from:

only the first plurality of class values when the robustness value indicates that the extracted metadata is valid and when the reliability value indicates that the first plurality of class values is reliable;

only the second plurality of class values when the robustness value indicates that the extracted metadata is invalid; and

the first plurality of class values and the second plurality of class values when the robustness value indicates that the extracted metadata is valid and the reliability value indicates that the first plurality of class values is unreliable.

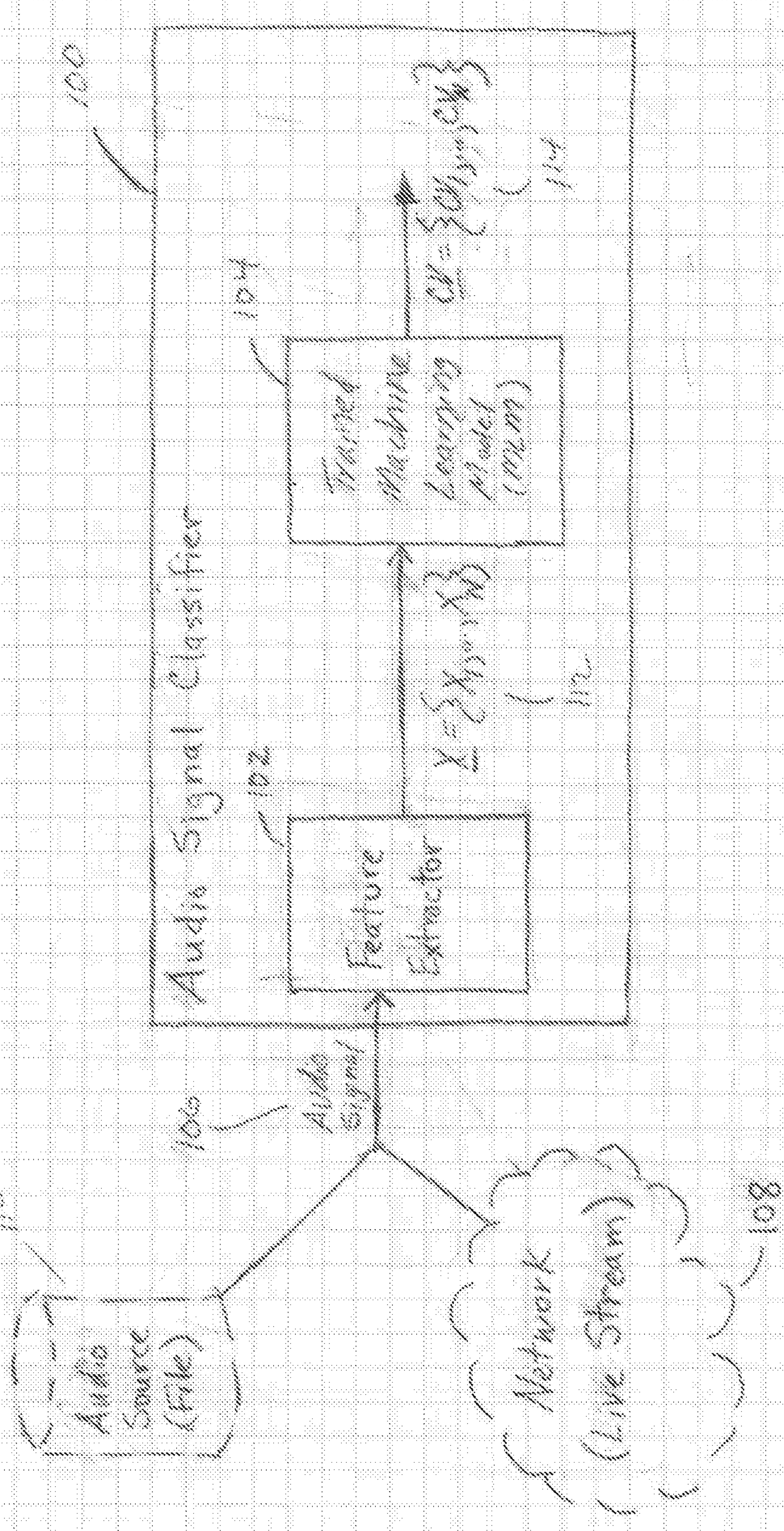
Audio Signal Classifier
100
Feature Extractor
102
Trained Machine Learning Model (MLM)
104
Audio Signal
106
Network (Live Stream)
108
Audio Source (File)
110
112
114

Fig. 1

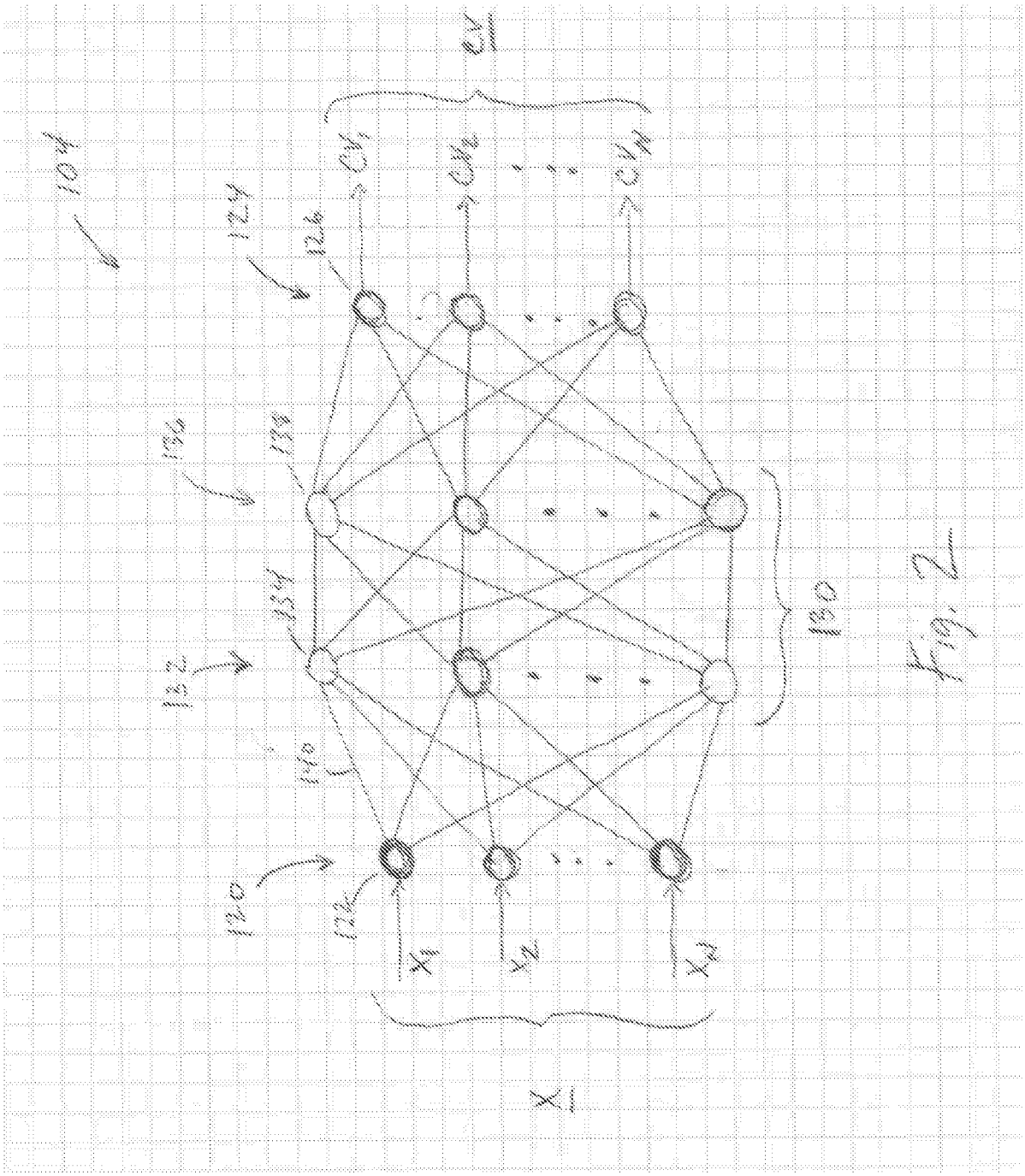
104
124
124a
cv_1
cv_2
cv_M
cv
116
118
154
122
110
120
120a
x_1
x_2
x_3
x_1
180

Fig. 2

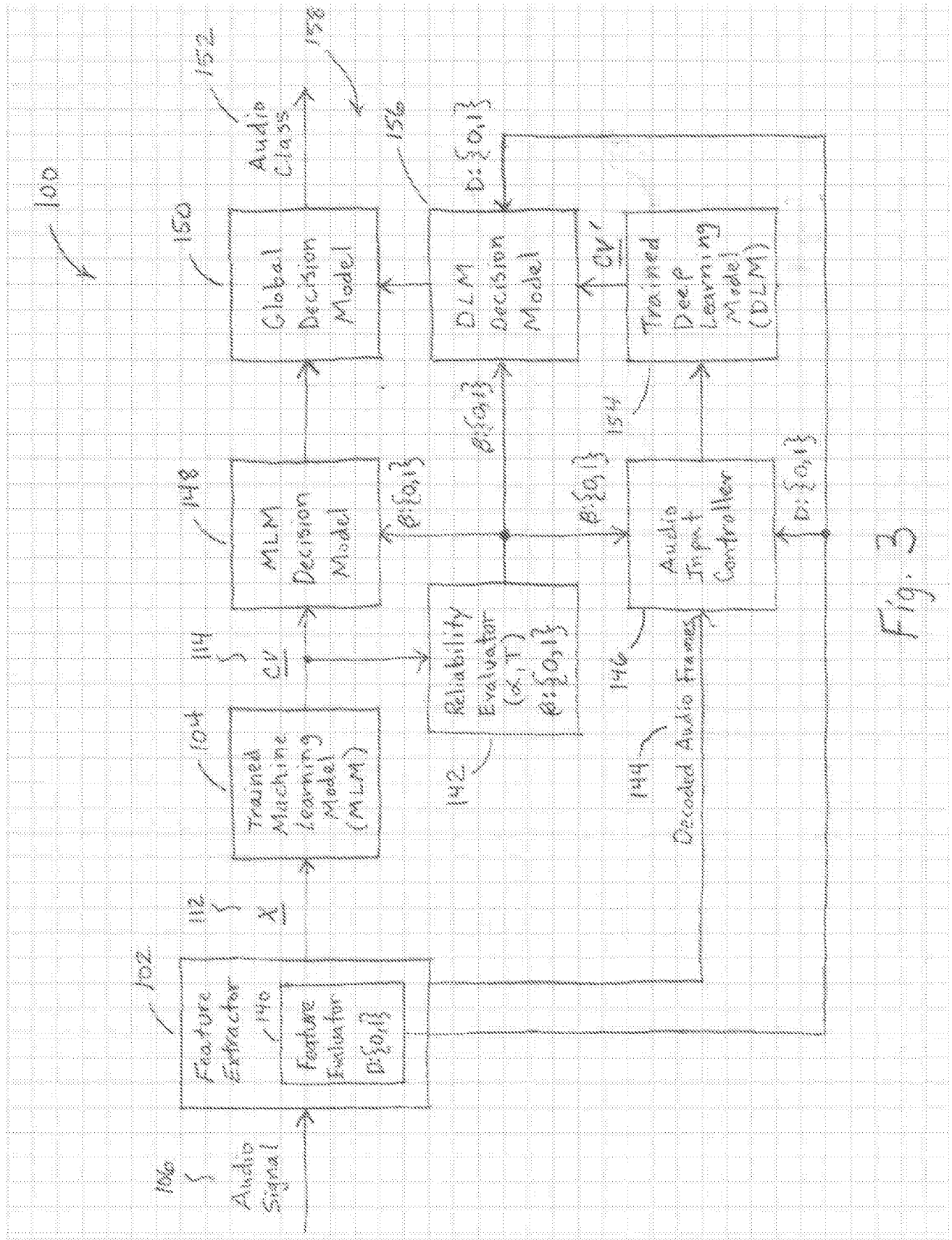
100
Audio Signal
106
Feature Extractor
102
Feature Evaluator
116
112
X
Trained Machine Learning Model (MLM)
104
CV
114
MLM Decision Model
148
Reliability Evaluator
142
Global Decision Model
150
Audio Class
152
158
DLM Decision Model
156
Trained Deep Learning Model (DLM)
154
Audio Info Controller
146
Decoded Audio Frames
144

Fig. 3

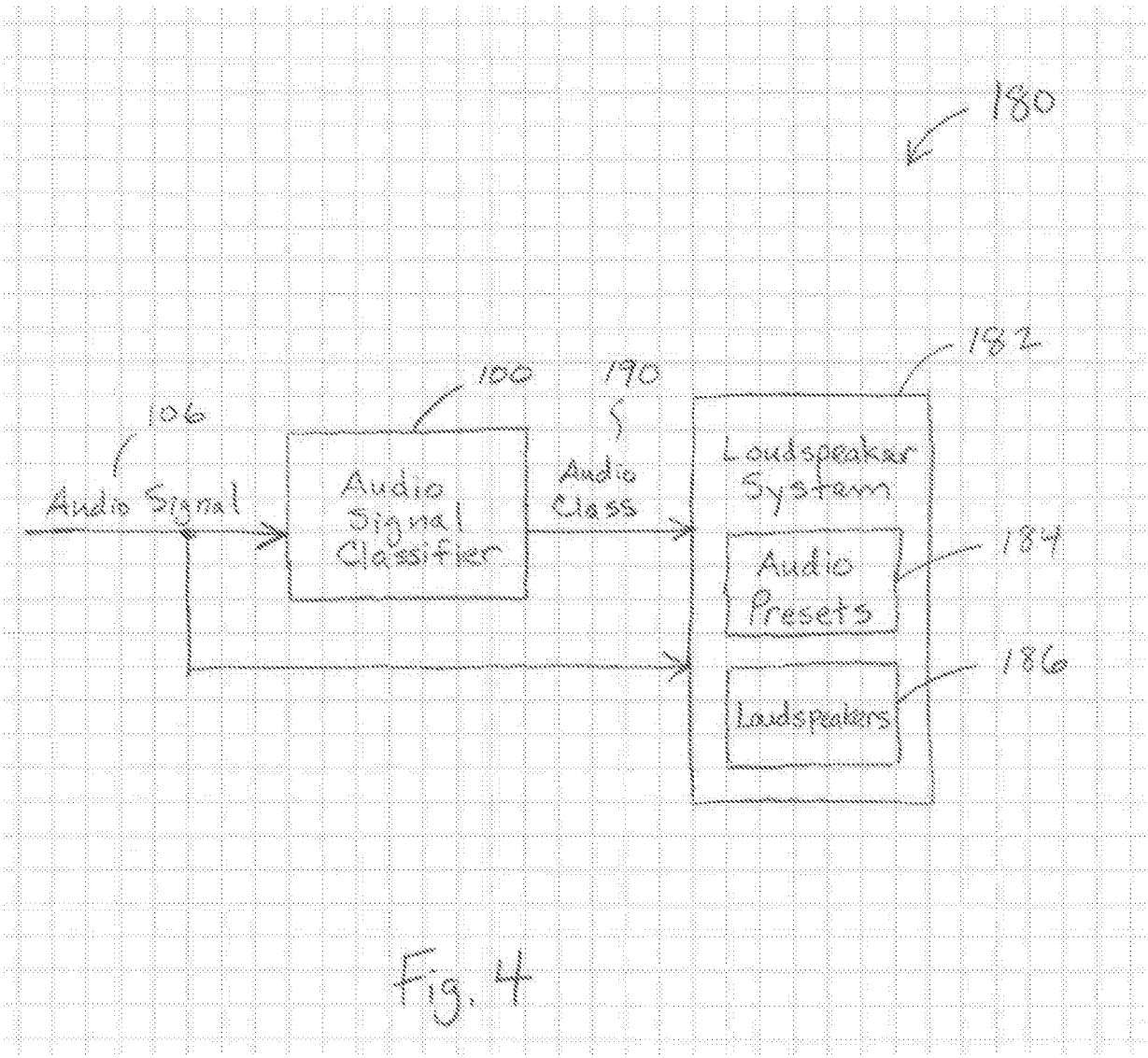
180
182
100
190
106
Audio Signal
Audio Signal Classifier
Audio Class
Loudspeaker System
Audio Presets
184
Loudspeakers
186

Fig. 4

	Lengths	Movies Unlimited Prices	Critic's Choice Prices
mean	129.6	14.09	13.00
mode	105 (3 times)	12.79, 14.99 (6 times)	12.71 (13 times)
minimum	81	5.99	5.82
first quartile	108	11.99	11.855
median	124.5	13.24	12.71
third quartile	139.5	15.99	12.735
maximum	200	29.99	25.46
range	119	21.00	19.64
standard deviation	23.3	4.02	3.08
interquartile range	31.5	4.00	0.88

Fig. 5

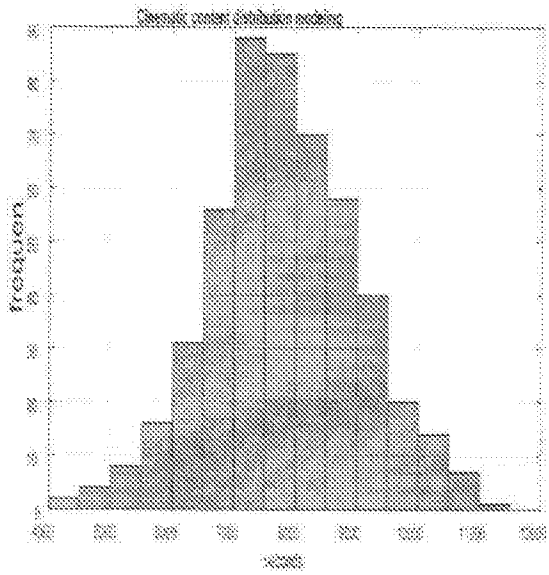

Fig. 6

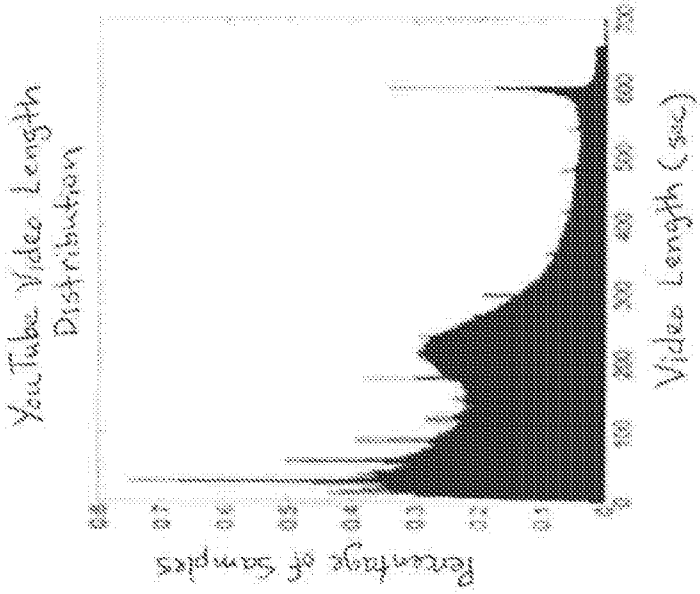
YouTube Video Length Distribution
Percentage of Samples
Video Length (sec)

Fig. 7A

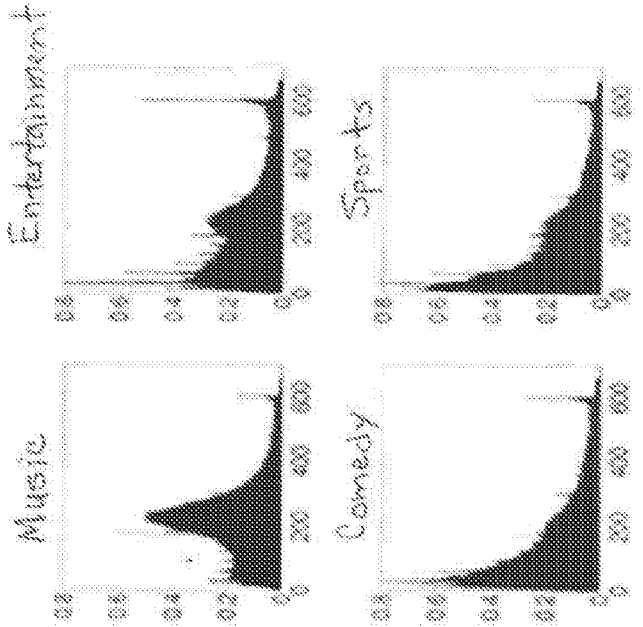
Entertainment
Sports
Music
Comedy

Fig. 7B

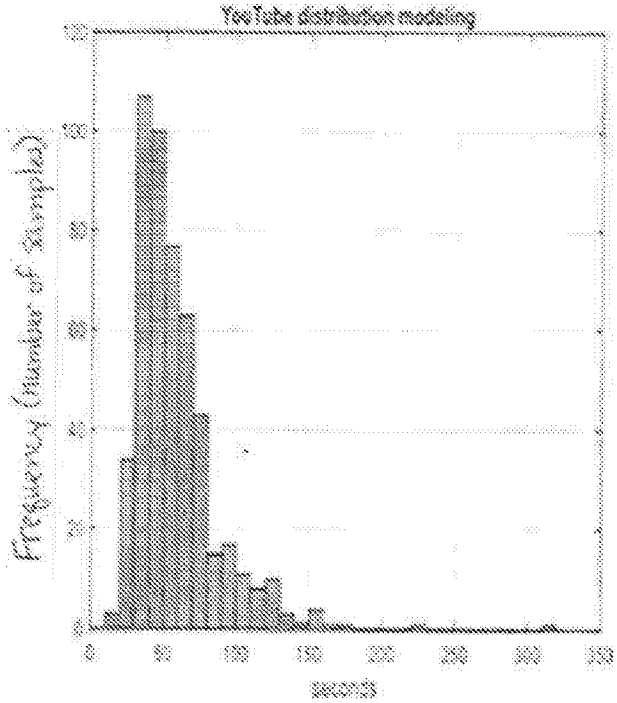
YouTube distribution modeling
Frequency (number of samples)
seconds

Fig. 8

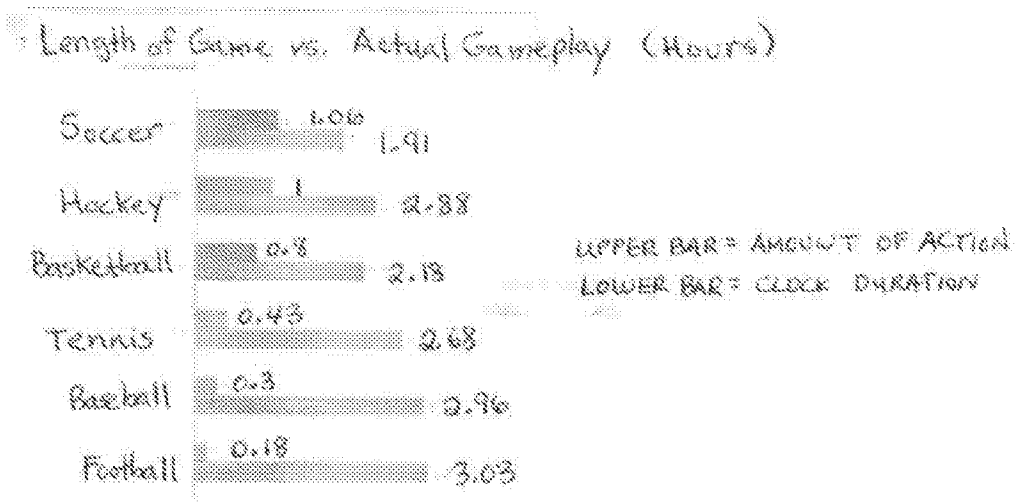
Length of Game vs. Actual Gameplay (Hours)
Soccer
1.06
1.91
Hockey
1
2.88
Basketball
0.8
2.18
Tennis
0.43
2.68
Baseball
0.3
3.96
Football
0.18
3.03
UPPER BAR = AMOUNT OF ACTION
LOWER BAR = CLOCK DURATION

Fig. 9

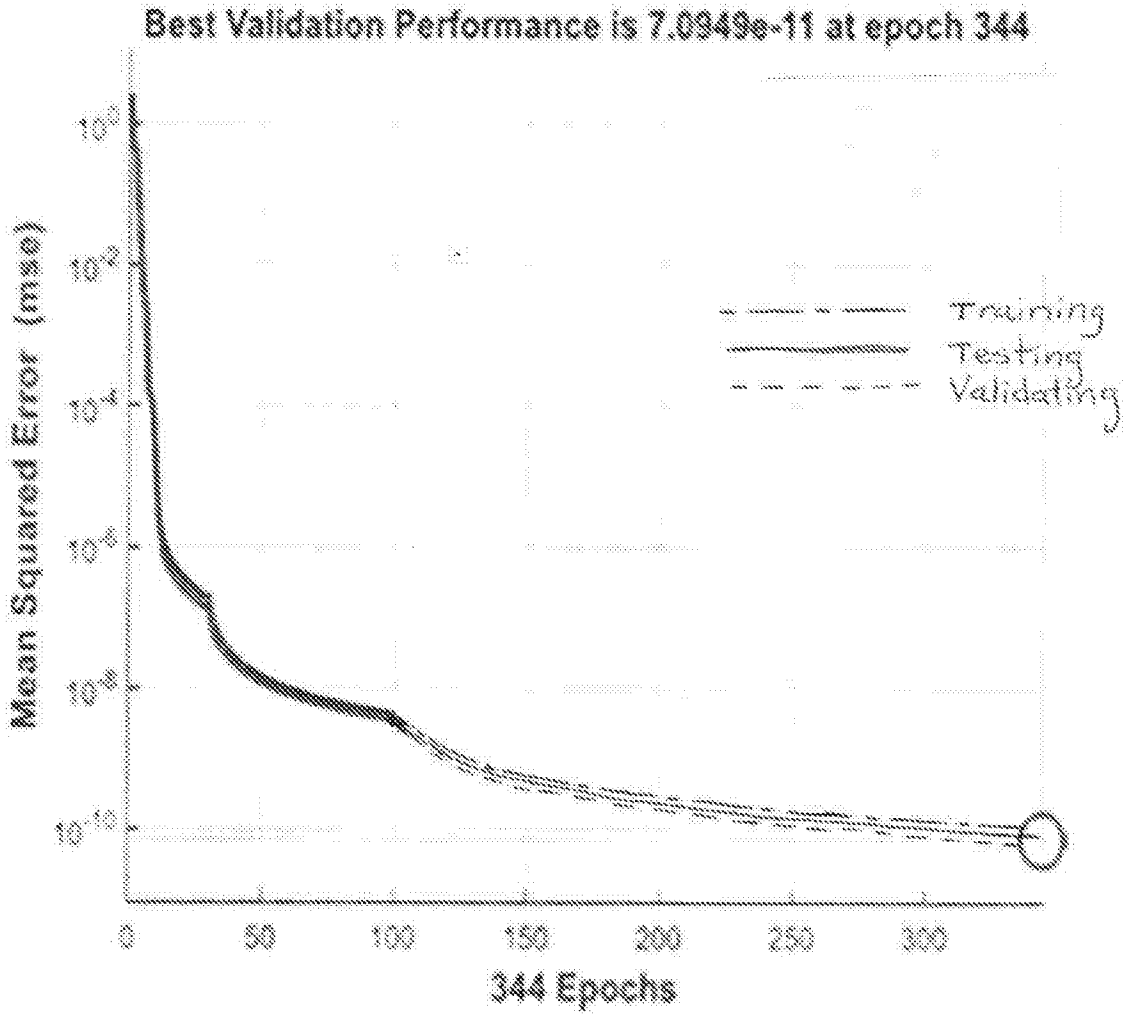
Best Validation Performance is 7.0949e-11 at epoch 344
Mean Squared Error (mse)
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
10^{-10}
Training
Testing
Validating
0
50
100
150
200
250
300
344 Epochs

Fig. 10

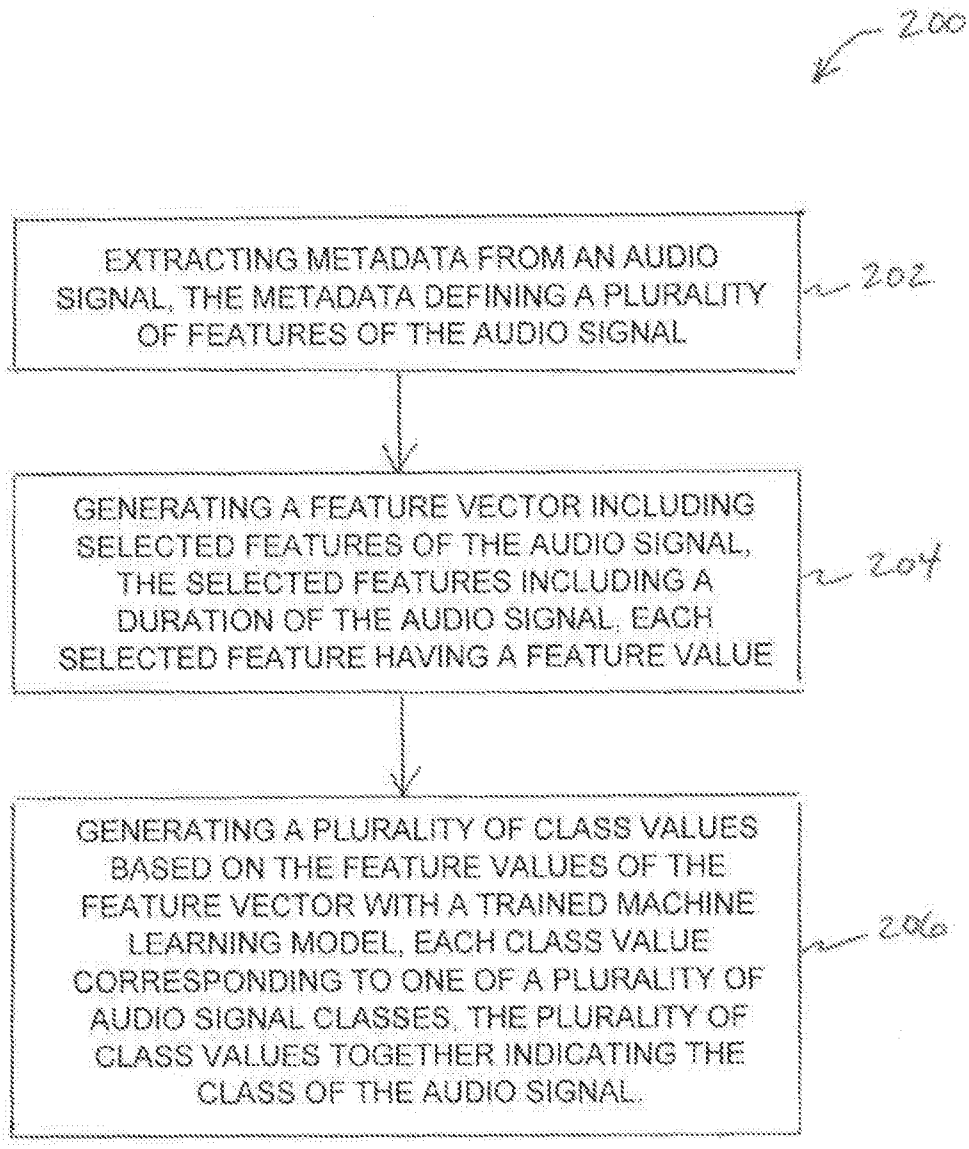
200
EXTRACTING METADATA FROM AN AUDIO SIGNAL, THE METADATA DEFINING A PLURALITY OF FEATURES OF THE AUDIO SIGNAL
202
GENERATING A FEATURE VECTOR INCLUDING SELECTED FEATURES OF THE AUDIO SIGNAL, THE SELECTED FEATURES INCLUDING A DURATION OF THE AUDIO SIGNAL, EACH SELECTED FEATURE HAVING A FEATURE VALUE
204
GENERATING A PLURALITY OF CLASS VALUES BASED ON THE FEATURE VALUES OF THE FEATURE VECTOR WITH A TRAINED MACHINE LEARNING MODEL, EACH CLASS VALUE CORRESPONDING TO ONE OF A PLURALITY OF AUDIO SIGNAL CLASSES, THE PLURALITY OF CLASS VALUES TOGETHER INDICATING THE CLASS OF THE AUDIO SIGNAL.
206

Fig. 11

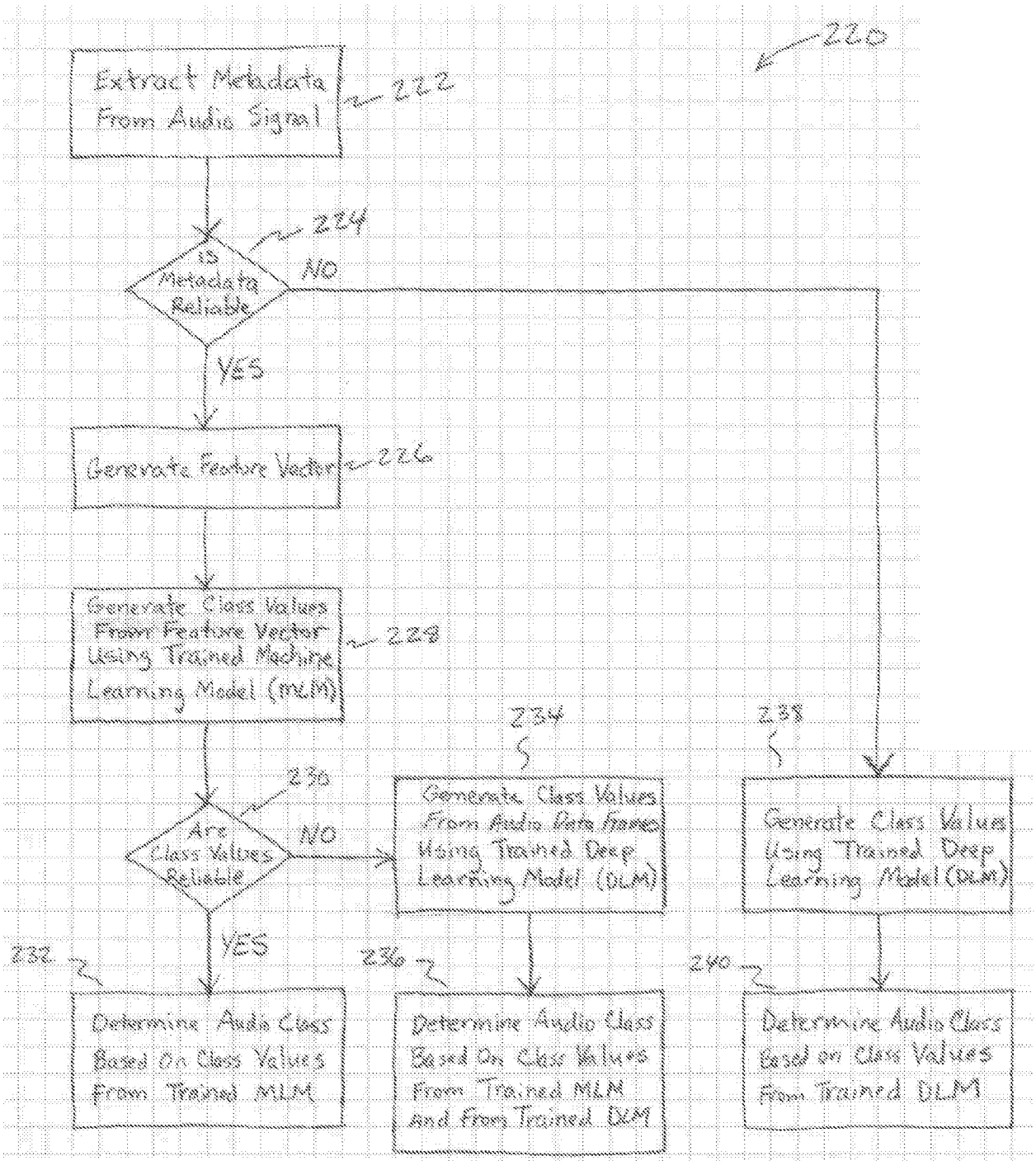
220
Extract Metadata From Audio Signal
222
Is Metadata Reliable
224
NO
YES
Generate Feature Vector
226
Generate Class Values From Feature Vector Using Trained Machine Learning Model (MLM)
228
Are Class Values Reliable
230
NO
YES
232
Determine Audio Class Based On Class Values From Trained MLM
234
Generate Class Values From Audio Data frames Using Trained Deep Learning Model (DLM)
236
Determine Audio Class Based On Class Values From Trained MLM and From Trained DLM
238
Generate Class Value Using Trained Deep Learning Model (DLM)
240
Determine Audio Class Based on Class Values From Trained DLM

Fig. 12

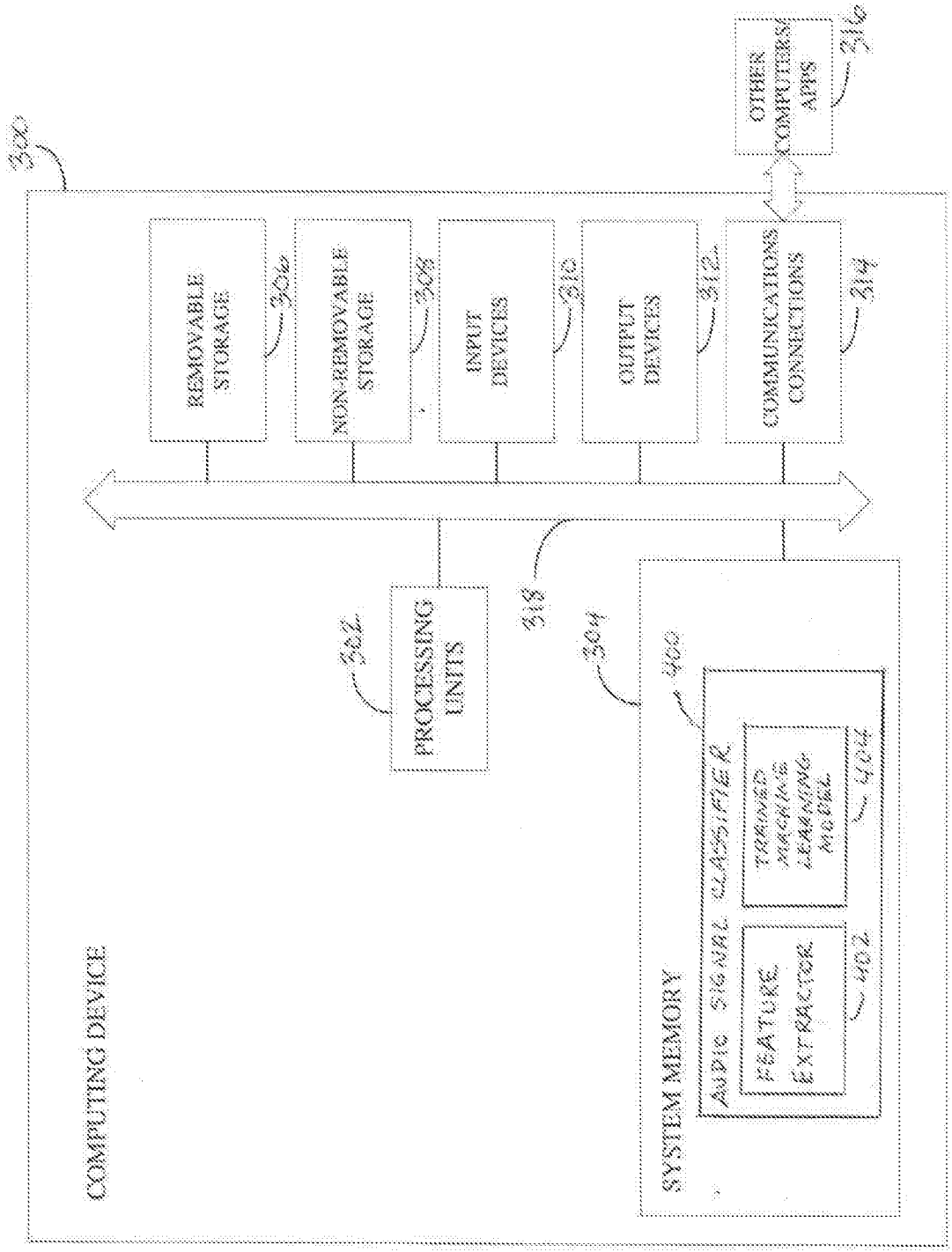
COMPUTING DEVICE
SYSTEM MEMORY
AUDIO SIGNAL CLASSIFIER
FEATURE EXTRACTOR
TRAINED MACHINE LEARNING MODEL
PROCESSING UNITS
REMOVABLE STORAGE
NON-REMOVABLE STORAGE
INPUT DEVICES
OUTPUT DEVICES
COMMUNICATIONS CONNECTIONS
OTHER COMPUTERS/ APPS
320
302
304
306
308
310
312
314
316
318
402
404

Fig. 13

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US 2017/030213

A. CLASSIFICATION OF SUBJECT MATTER

G06F 15/18 *(2006.01)*
H04R 1/20 *(2006.01)*

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)

G06F 15/00, 15/18, G06N 5/00, H04R 1/00, 1/20, 1/22, G10L 25/00, 25/78, 25/84

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

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

DWPI, ESP@CENET, K-PION, PAJ, SIPO, PatSearch, RUPTO, USPTO, WIPO,

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2009/0012638 A1 (XIA LOU) 08.01.2009, paragraphs [0010], [0022], [0032]-[0038], [0044]	1-15
A	US 2016/0302014 A1 (KELLY FITZ et al.) 13.10.2016, paragraphs [0007], [0026]	1-15

☐ 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

"E" earlier document 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	Date of mailing of the international search report
11 January 2018 (11.01.2018)	18 January 2018 (18.01.2018)

Name and mailing address of the ISA/RU:
Federal Institute of Industrial Property,
Berezhkovskaya nab., 30-1, Moscow, G-59,
GSP-3, Russia, 125993
Facsimile No: (8-495) 531-63-18, (8-499) 243-33-37

Authorized officer

T.Kiseleva

Telephone No. 495 531 65 15