



(51) International Patent Classification:

G10K 11/175 (2006.01)

(21) International Application Number:

PCT/SG2023/050289

(22) International Filing Date:

26 April 2023 (26.04.2023)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

10202204451S 27 April 2022 (27.04.2022) SG

(71) Applicant: **NANYANG TECHNOLOGICAL UNIVERSITY** [SG/SG]; 50 Nanyang Avenue, Singapore 639798 (SG).

(72) Inventors: **OOI, Wen Rui Kenneth**; c/o Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798 (SG). **WATCHARASUPAT, Karn**; c/o Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798 (SG). **LAM, Bhan**; c/o Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798 (SG). **ONG, Zhen Ting**; c/o Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798 (SG). **WONG, Trevor Martens Zhi Ming**; c/o Nanyang Technological

University, 50 Nanyang Avenue, Singapore 639798 (SG). **GAN, Woon Seng**; c/o Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798 (SG).

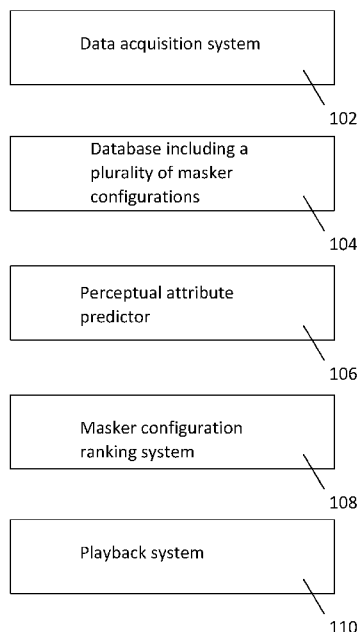
(74) Agent: **CHINA SINDA INTELLECTUAL PROPERTY PTE. LTD.**; 47 Scotts Road, #13-03, Goldbell Towers, Singapore 228233 (SG).

(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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT,

(54) Title: SOUNDSCAPE AUGMENTATION SYSTEM AND METHOD OF FORMING THE SAME

FIG. 1



(57) Abstract: Various embodiments may provide a soundscape augmentation system. The soundscape augmentation system may include a data acquisition system configured to provide ambient soundscape data. The soundscape augmentation system may also include a database including a plurality of masker configurations. The soundscape augmentation system may further include a perceptual attribute predictor coupled to the data acquisition system and the database, the perceptual attribute predictor configured to generate predictions representing perception on one or more pre-defined perceptual attribute scales for each masker configuration of the plurality of masker configurations based on the ambient soundscape data. The soundscape augmentation system may additionally include a masker configuration ranking system configured to determine one or more optimal masker configurations based on the predictions generated by the perceptual attribute predictor. The soundscape augmentation system may also include a playback system configured to play back or reproduce the one or more optimal masker configurations.

LU, LV, MC, ME, 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).

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

SOUNDSCAPE AUGMENTATION SYSTEM AND METHOD OF FORMING THE SAME

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims the benefit of priority of Singapore application No. 10202204451S filed April 27, 2022, the contents of it being hereby incorporated by reference in its entirety for all purposes.

TECHNICAL FIELD

[0002] Various embodiments of this disclosure may relate to a soundscape augmentation system. Various embodiments of this disclosure may relate to a method of forming a soundscape augmentation system.

BACKGROUND

[0003] The World Health Organization (WHO) has labelled exposure to noise as the next to worst environmental pollutant after air pollution. Noise exposure has also been likened to second-hand smoke. The urgent call to action to mitigate noise exposure arises from increasing evidence of noise-induced health effects such as increased risk of ischemic heart disease incidence, annoyance and sleep disturbance. The WHO also highlights the impact of environmental noise exposure to mental health and well-being, with new evidence showing a harmful effect on measures of depression and anxiety.

[0004] There is an emerging paradigm shift from noise management to soundscape management. Soundscape is defined in ISO 12913-1 as an “acoustic environment as perceived or experienced and/or understood by a person or people, in context”. As opposed to noise management, the soundscape management framework perceives sound as a resource rather than

a waste; focuses on sounds of preference rather than sounds of discomfort; and manages masking unwanted with wanted sounds as well as reducing unwanted sounds rather than just reducing sound levels. Hence, there are soundscape intervention techniques based on augmentation or introduction of “masking” sounds into the acoustic environment to improve overall perception of acoustic comfort. Commonly, such interventions involve the augmentation with natural sounds (e.g., via loudspeakers) and have been trialled in outdoor recreational spaces, and indoors, such as in nursing homes, to improve acoustic comfort. Augmentation of soundscapes with wanted sounds is akin to sound masking systems, which are commonly employed in office environments to reduce distractions. The key difference between sound masking systems and the soundscape-based augmentation systems is that masking is based on objective metrics, whereas soundscape augmentation is based largely on subjective metrics rooted in context. The context “includes the interrelationships between person and activity and place, in space and time” and is further detailed with examples in ISO 12913-1. It is also worth noting that soundscapes are not limited to real-world environments, but can also encompass virtual environments and even recollection from memory.

SUMMARY

[0005] Various embodiments may provide a soundscape augmentation system. The soundscape augmentation system may include a data acquisition system configured to provide ambient soundscape data. The soundscape augmentation system may also include a database including a plurality of masker configurations. The soundscape augmentation system may further include a perceptual attribute predictor coupled to the data acquisition system and the database, the perceptual attribute predictor configured to generate predictions representing perception on one or more pre-defined perceptual attribute scales for each masker configuration of the plurality of masker configurations based on the ambient soundscape data. The soundscape

augmentation system may additionally include a masker configuration ranking system configured to determine one or more optimal masker configurations based on the predictions generated by the perceptual attribute predictor. The soundscape augmentation system may also include a playback system configured to play back or reproduce the one or more optimal masker configurations.

[0006] Various embodiments may relate to a method of forming a soundscape augmentation system. The method may include providing a data acquisition system configured to provide ambient soundscape data. The method may also include providing a database including a plurality of masker configurations. The method may further include coupling a perceptual attribute predictor to the data acquisition system and the database, the perceptual attribute predictor configured to generate predictions representing perception on one or more pre-defined perceptual attribute scales for each masker configuration of the plurality of masker configurations based on the ambient soundscape data. The method may additionally include providing a masker configuration ranking system configured to determine one or more optimal masker configurations based on the predictions generated by the perceptual attribute predictor. The method may further include providing a playback system configured to play back or reproduce the one or more optimal masker configurations.

BRIEF DESCRIPTION OF THE DRAWINGS

[0007] In the drawings, like reference characters generally refer to the same parts throughout the different views. The drawings are not necessarily drawn to scale, emphasis instead generally being placed upon illustrating the principles of various embodiments. In the following description, various embodiments of the invention are described with reference to the following drawings.

FIG. 1 shows a schematic of a soundscape augmentation system according to various embodiments.

FIG. 2 shows a schematic illustrating a method of forming a soundscape augmentation system according to various embodiments.

FIG. 3 shows a schematic of a soundscape augmentation system according to various embodiments.

FIG. 4 shows the two-dimensional circumplex octant model of affective quality attributes as presented in ISO12913-3 in which predictions from the perceptual attribute predictor according to various embodiments may be based on.

FIG. 5 is a schematic of the perceptual attribute predictor according to various embodiments.

FIG. 6 is a schematic of the perceptual attribute predictor according to various embodiments.

FIG. 7 is a schematic of the perceptual attribute predictor according to various embodiments.

FIG. 8 shows a training and inference schema of an automatic masker selection system (AMSS) according to various embodiments.

FIG. 9 shows (a) the base convolutional recurrent neural network (CRNN) architecture used according to various embodiments; (b) one possible implementation of the feature mapping block in (a) according to various embodiments; and (c) other possible implementations of the feature mapping block in (a) according to various other embodiments.

FIG. 10 is a table showing the mean fold mean squared errors (MSEs) of the probabilistic perceptual attribute predictor (PPAP) ($\pm$ standard deviation) over the 10 runs tested for each setting (above: cross-validation set, below: test set) according to various embodiments.

FIG. 11 illustrates masker selection using the 50 models (dot-product variant) on the test set, with a naïve maximum μ_k selection scheme (left) according to various embodiments and a random sampling scheme to encourage masker exploration (right) according to various embodiments.

FIG. 12A shows a schematic of the probabilistic perceptual attribute predictor (PPAP) according to various embodiments.

FIG. 12B illustrates three feature augmentation methods according to Equations (8), (9) and (10) in various embodiments.

FIG. 13 shows an algorithm to reduce total runtime according to various embodiments.

FIG. 14 shows (above) a plot of mean squared error (MSE) as a function of attention block type showing violin plots of the validation MSE of the pleasantness prediction for each attention block type and feature augmentation method according to various embodiments; and (below) a plot of mean absolute error (MAE) as a function of attention block type and feature augmentation method showing violin plots of the validation MAE of the pleasantness prediction according to various embodiments.

FIG. 15A is a plot of ISO Pleasantness as a function of log gain illustrating gain interpolation using an additive attention (AA) with CONV augmentation (seed 5, validation fold 2) for masker class “water” according to various embodiments. FIG. 15B is a plot of ISO Pleasantness as a function of log gain illustrating gain interpolation using an additive attention (AA) with CONV augmentation (seed 5, validation fold 2) for masker class “traffic” according to various embodiments.

FIG. 15C is a plot of ISO Pleasantness as a function of log gain illustrating gain interpolation using an additive attention (AA) with CONV augmentation (seed 5, validation fold 2) for masker class “bird” according to various embodiments.

FIG. 15D is a plot of ISO Pleasantness as a function of log gain illustrating gain interpolation using an additive attention (AA) with CONV augmentation (seed 5, validation fold 2) for masker class “construction” according to various embodiments.

FIG. 15E is a plot of ISO Pleasantness as a function of log gain illustrating gain interpolation using an additive attention (AA) with CONV augmentation (seed 5, validation fold 2) for masker class “silence” according to various embodiments.

DESCRIPTION

[0008] The following detailed description refers to the accompanying drawings that show, by way of illustration, specific details and embodiments in which the invention may be practised. These embodiments are described in sufficient detail to enable those skilled in the art to practise the invention. Other embodiments may be utilized and structural, logical, and electrical changes may be made without departing from the scope of the invention. The various embodiments are not necessarily mutually exclusive, as some embodiments can be combined with one or more other embodiments to form new embodiments.

[0009] Features that are described in the context of an embodiment may correspondingly be applicable to the same or similar features in the other embodiments. Features that are described in the context of an embodiment may correspondingly be applicable to the other embodiments, even if not explicitly described in these other embodiments. Furthermore, additions and/or combinations and/or alternatives as described for a feature in the context of an embodiment may correspondingly be applicable to the same or similar feature in the other embodiments.

[0010] In the context of various embodiments, the articles “a”, “an” and “the” as used with regard to a feature or element include a reference to one or more of the features or elements.

[0011] In the context of various embodiments, the term “about” or “approximately” as applied to a numeric value encompasses the exact value and a reasonable variance, e.g. within 10% of the specified value.

[0012] As used herein, the term “and/or” includes any and all combinations of one or more of the associated listed items.

[0013] By “comprising” it is meant including, but not limited to, whatever follows the word “comprising”. Thus, use of the term “comprising” indicates that the listed elements are required or mandatory, but that other elements are optional and may or may not be present.

[0014] By “consisting of” is meant including, and limited to, whatever follows the phrase “consisting of”. Thus, the phrase “consisting of” indicates that the listed elements are required or mandatory, and that no other elements may be present.

[0015] Embodiments described in the context of one of the soundscape augmentation systems are analogously valid for the other soundscape augmentation systems. Similarly, embodiments described in the context of a method are analogously valid for a soundscape augmentation system, and vice versa.

[0016] The current state-of-the-art augmentation or masking interventions involve the static playback of sounds through a loudspeaker, wherein the sound levels and tracks have to be manually selected. Various embodiments may utilize a first-of-its-kind artificial intelligence model to automatically select a soundtrack that yields the highest (or lowest) value on any arbitrary subjective metric (e.g., perceived loudness, circumplex octant scale from ISO 12913-3 – pleasantness, annoyance, tranquillity, calmness, vibrancy, eventfulness, etc.) at the most appropriate sound level. The artificial intelligence (AI) model is trained on a large dataset of human subjective responses of the local Singapore population collected by the research team at Nanyang Technological University (NTU) based on ISO 12913-2 standards. The ISO 12913 series of standards represents a paradigm shift in sound environment management and details perceptually-based methods to holistically assess and analyse the sound environment. In addition, the AI model doubles up as a prediction tool to automatically assess perceptual qualities of the sound environment without the use of tedious and labour-intensive surveys or questionnaires as required in ISO 12913-2.

[0017] FIG. 1 shows a schematic of a soundscape augmentation system according to various embodiments. The soundscape augmentation system may include a data acquisition system 102 configured to provide ambient soundscape data. The soundscape augmentation system may also include a database 104 including a plurality of masker configurations. The soundscape augmentation system may further include a perceptual attribute predictor 106 coupled to the data acquisition system 102 and the database 104, the perceptual attribute predictor 106 configured to generate predictions representing perception on one or more pre-defined perceptual attribute scales for each masker configuration of the plurality of masker configurations based on the ambient soundscape data. The soundscape augmentation system may additionally include a masker configuration ranking system 108 configured to determine one or more optimal masker configurations based on the predictions generated by the perceptual attribute predictor 106. The soundscape augmentation system may also include a playback system 110 configured to play back or reproduce the one or more optimal masker configurations.

[0018] In other words, the soundscape augmentation system may include a perceptual attribute predictor 106 which may be connected to a data acquisition system 102 and a database 104. The soundscape augmentation system may also include a masker configuration ranking system 108 connected to the perceptual attribute predictor 106 and a playback system 110 connected to the masker configuration ranking system 108.

[0019] For avoidance of doubt, FIG. 1 is intended to illustrate some features of a soundscape augmentation system according to various embodiments, and is not intended to limit, for instance, the size, shape, orientation, arrangement etc. of the various components.

[0020] In various embodiments, the data acquisition system 102 may be also configured to provide demographic or contextual data. The perceptual attribute predictor 106 may be configured to generate the predictions also based on the demographic or contextual data. The

demographic or contextual data may be received, measured or sensed by the data acquisition system 102 directly or received from external sources, e.g. devices such as portable health monitoring devices or storage devices, as described in more detail below.

[0021] The demographic or contextual data may include data related to environmental parameters, demographic data of one or more listeners, psychological data of the one or more listeners, physiological data of the one or more listeners, or any combination thereof.

[0022] In this regard, the one or more listeners may include real and/or hypothetical listeners.

[0023] In various embodiments, the environmental data related to environmental parameters may be or include data related to location, data related to the visual environment, number of people nearby, air temperature, humidity, wind speed, or any combination thereof. The environmental parameters may be non-acoustic environmental parameters. For instance, the data acquisition system 102 may include or may be coupled to a Global Positioning System (GPS) sensor to determine the data related to location. The data acquisition system 102 may include or may be coupled to a camera to determine the data related to number of people nearby and/or to obtain visual information from the environment (e.g. determine presence of specific types of furniture, pets etc., determine movement of people, determine the percentage of greenery or water present etc.). The data acquisition system 102 may include or may be coupled to a thermometer to determine the data related to air temperature. The data acquisition system 102 may include or may be coupled to a hygrometer to determine the data related to humidity. The data acquisition system 102 may include or may be coupled to an anemometer to determine the data related to wind speed. In the present context, the term “coupled” may refer to being connected either directly or indirectly. The connection may be established, for instance, via wired or wireless means.

[0024] In various embodiments, the demographic data of the one or more listeners may be or include data related to age, gender, occupation, or any combination thereof. In various embodiments, the demographic data of the one or more listeners provided by the data acquisition system 102 may be obtained by the data acquisition system 102 from external sources. In various embodiments, the demographic data of the one or more listeners may be provided via a digital form attached to the data acquisition system 102 or as pre-loaded data on a storage device coupled to the data acquisition system 102 (e.g. directly, or via a cloud-based server). In various embodiments, the demographic data of the one or more listeners may be provided to the data acquisition system 102 manually, e.g. keyed in to the data acquisition system 102 manually, or sent to the data acquisition system 102 from a remote system upon manual instructions provided by a user. In various embodiments, the data acquisition system 102 may be configured to obtain the demographic data of the one or more listeners automatically, e.g. from a remote system via wired or wireless means.

[0025] In various embodiments, the psychological data of the one or more listeners may be or may include data related to noise sensitivity (e.g. using the Weinstein Noise Sensitivity Scale), perceived stress (e.g. using Cohen's Perceived Stress Scale), well-being index scores (e.g. using the WHO-5 Well-Being Index), or any combination thereof. In various embodiments, the psychological data of the one or more listeners provided by the data acquisition system 102 may be obtained by the data acquisition system 102 from external sources. In various embodiments, the psychological data of the one or more listeners may be provided via a digital form attached to the data acquisition system 102 or as pre-loaded data on a storage device coupled to the data acquisition system 102 (e.g. directly, or via a cloud-based server). In various embodiments, the psychological data of the one or more listeners may be provided to the data acquisition system 102 manually, e.g. keyed in to the data acquisition system 102 manually, or sent to the data acquisition system 102 from a remote system upon

manual instructions provided by a user. In various embodiments, the data acquisition system 102 may be configured to obtain the psychological data of the one or more listeners automatically, e.g. from a remote system via wired or wireless means.

[0026] In various embodiments, the physiological data of the one or more listeners may be or include data related to heart rate, blood pressure, body temperature, or any combination thereof. For instance, the data acquisition system 102 may include or be coupled (e.g. directly or via a cloud-based server) to one or more portable health monitoring devices, which are configured to determine or measure the physiological data of the one or more listeners. The data acquisition system 102 may be configured to obtain the physiological data of the one or more listeners from one or more portable health monitoring devices via wired or wireless means.

[0027] In various embodiments, the perceptual attribute predictor 106 may be configured to generate predictions also based on one or more masker gain inputs. The one or more masker gain inputs may, for instance, be one or more digital gain levels, or one or more masker gain waveforms.

[0028] In various embodiments, the one or more pre-defined perceptual attribute scales may include pleasantness, vibrancy, eventfulness, calmness, perceived loudness, sound quality, sharpness, roughness, or any combination thereof. In various embodiments, the one or more pre-defined perceptual attribute scales may be selected from the affective quality attributes as defined in ISO12913-3.

[0029] In various embodiments, the predictions generated may or may not be of single numerical values. In various embodiments, the predictions generated may be non-deterministic. In various other embodiments, the predictions generated may be deterministic. In various embodiments, the predictions may be a probability distribution, random values extracted from the probability distribution, a vector or a multivariate representation of the one or more pre-

defined perceptual attributes. In various embodiments, the predictions generated may be a predicted attribute distribution.

[0030] In various embodiments, the perceptual attribute predictor 106 may be or may include a probabilistic perceptual attribute predictor. In various embodiments, the perceptual attribute predictor 106 may be configured to generate the predictions by mixing, combining or adding the ambient soundscape data and each masker configuration. In various embodiments, the ambient soundscape data and each masker configuration may be added, mixed or combined, e.g. in an inference engine, to form augmented soundscape data before the augmented soundscape data is being inputted into the perceptual attribute predictor 106. A masker waveform of each masker configuration may be weighted (i.e. multiplied) with a masker gain input, e.g. a masker gain waveform, to generate a weighted masker waveform. The weighted masker waveform may be combined with (i.e. added to) an ambient soundscape waveform of the ambient soundscape data to generate an augmented soundscape waveform. The augmented soundscape waveform may be received by a prediction block of the perceptual attribute predictor 106 to generate the predictions.

[0031] In various embodiments, the perceptual attribute predictor 106 may include a soundscape feature extractor configured to extract features of the ambient soundscape data. The perceptual attribute predictor 106 may include a masker feature extractor configured to extract features of each masker configuration. The perceptual attribute predictor 106 may include a feature level augmentor configured to generate one or more augmented soundscape features based on the extracted features of the ambient soundscape data, the extracted features of each masker configuration and a masker gain input, e.g. a digital gain level. The perceptual attribute predictor 106 may include a prediction block configured to generate the predictions based on the one or more augmented soundscape features, the extracted features of the ambient soundscape data and the extracted features of each masker configuration. The perceptual

attribute predictor 106 may be configured to generate the predictions based on the extracted features of the ambient soundscape data and the extracted features of each masker configuration.

[0032] In various embodiments, the perceptual attribute predictor 106 may include an objective feature block configured to extract objective features from the ambient soundscape data and each masker configuration. The perceptual attribute predictor 106 may be further configured to extract the objective features also based on a masker gain input. The perceptual attribute predictor 106 may include a subjective feature block configured to extract subjective features from the demographic or contextual data. The subjective feature block may also be configured to process the extracted subjective features and the extracted objective features to generate the predictions. The perceptual attribute predictor 106 may be configured to generate the predictions based on the extracted objective features and the extracted subjective features.

[0033] In various embodiments, the perceptual attribute predictor 106 may include linear regression models configured to generate the predictions based on acoustic or psychoacoustic parameters computed based on the ambient soundscape data and each masker configuration.

[0034] In various embodiments, the perceptual attribute predictor 106 may include one or more deep neural networks configured to generate the predictions based on raw audio, spectrogram representations computed based on the ambient soundscape data and each masker configuration, or a combination of the raw audio and the spectrogram representations.

[0035] In various embodiments, the perceptual attribute predictor 106 may include a composite network as described herein, or any combination of networks and/or models as described herein.

[0036] In various embodiments, the ambient soundscape data may be received, recorded, measured or sensed by the data acquisition system 102 directly or received from external sources. In various embodiments, the data acquisition system 102 may be configured to provide

or generate the ambient soundscape data based on inputs from recording devices, receiving devices, and/or storage devices. The data acquisition system 102 may be configured to provide or generate the ambient soundscape data based on inputs from one or more microphones, one or more antennas, a storage medium, or any combination thereof. In various embodiments, the data acquisition system 102 may be configured to provide or generate the ambient soundscape data by directly recording or sensing the ambient soundscape of an environment.

[0037] In various embodiments, the plurality of masker configurations may include one or more audio tracks of recorded or synthesized sounds, one or more audio tracks of silence, or one or more audio tracks derived from the one or more audio tracks of recorded or synthesized sounds and the one or more audio tracks of silence.

[0038] In various embodiments, the playback system 110 may include one or more speakers, one or more virtual or augmented reality headsets, one or more headphones, one or more earbuds, or any combination thereof.

[0039] In various embodiments, various components of the soundscape augmentation system may be implemented using one or more computing, processing or electronic devices or systems. In one example, the data acquisition system 102, the database 104, the perceptual attribute predictor 106, the masker configuration ranking system 108, the playback system 110 may be implemented in a single computing or processing device or system including a microphone and speakers, such as a physical or cloud-based server, a personal computer system or a mobile device. In another example, the data acquisition system 102 may be implemented via a mobile device (with a microphone), while the database 104, the perceptual attribute predictor 106, the masker configuration ranking system 108, and the playback system 110 may be implemented via a personal computer system with speakers in wireless communication (e.g. WiFi or Bluetooth) directly or indirectly with the mobile device. In yet another example, the data acquisition system 102 may be implemented with a laptop connected remotely to a

microphone and environmental sensors. The database 104, the perceptual attribute predictor 106, and the masker configuration ranking system 108 may be implemented using another computing device in communication to the laptop via a server. The playback system 110 may be via headphones in wireless communication with the computer device.

[0040] FIG. 2 shows a schematic illustrating a method of forming a soundscape augmentation system according to various embodiments. The method may include, in 202, providing a data acquisition system configured to provide ambient soundscape data. The method may also include, in 204, providing a database including a plurality of masker configurations. The method may further include, in 206, coupling a perceptual attribute predictor to the data acquisition system and the database, the perceptual attribute predictor configured to generate predictions representing perception on one or more pre-defined perceptual attribute scales for each masker configuration of the plurality of masker configurations based on the ambient soundscape data. The method may additionally include, in 208, providing a masker configuration ranking system configured to determine one or more optimal masker configurations based on the predictions generated by the perceptual attribute predictor. The method may further include, in 210, providing a playback system configured to play back or reproduce the one or more optimal masker configurations.

[0041] In other words, the method may include coupling a perceptual attribute predictor to a data acquisition system and a database containing a plurality of masker configurations. The method may also include coupling a masker configuration ranking system to the perceptual attribute predictor, and a playback system to the masker configuration ranking system.

[0042] For avoidance of doubt, FIG. 2 is not intended to limit the sequence of the various steps. For instance, step 202 may occur before, after or at the same time as step 204.

[0043] In various embodiments, the data acquisition system may be also configured to provide demographic or contextual data. The perceptual attribute predictor may be configured to generate the predictions also based on the demographic or contextual data.

[0044] In various embodiments, the demographic or contextual data may include data related to environmental parameters, demographic data of one or more listeners, psychological data of the one or more listeners, physiological data of the one or more listeners, or any combination thereof.

[0045] In various embodiments, the data related to environmental parameters may include data related to location, data related to the visual environment, number of people nearby, air temperature, humidity, wind speed, or any combination thereof.

[0046] In various embodiments, the demographic data of the one or more listeners may include data related to age, gender, occupation, or any combination thereof.

[0047] In various embodiments, the psychological data of the one or more listeners may include data related to noise sensitivity, perceived stress, well-being index scores, or any combination thereof.

[0048] In various embodiments, the physiological data of the one or more listeners may include data related to heart rate, blood pressure, body temperature, or any combination thereof.

[0049] In various embodiments, the perceptual attribute predictor may be configured to generate predictions based on one or more masker gain inputs.

[0050] In various embodiments, the one or more pre-defined perceptual attribute scales may include pleasantness, vibrancy, eventfulness, calmness, perceived loudness, sound quality, sharpness, roughness, or any combination thereof.

[0051] In various embodiments, the perceptual attribute predictor may be configured to generate the predictions by combining the ambient soundscape data and each masker configuration.

[0052] In various embodiments, the perceptual attribute predictor may include a soundscape feature extractor configured to extract features of the ambient soundscape data. The perceptual attribute predictor may include a masker feature extractor configured to extract features of each masker configuration. The perceptual attribute predictor may be configured to generate the predictions based on the extracted features of the ambient soundscape data and the extracted features of each masker configuration.

[0053] In various embodiments, the perceptual attribute predictor may include an objective feature block configured to extract objective features from the ambient soundscape data and each masker configuration. The perceptual attribute predictor may include a subjective feature block configured to extract subjective features from the demographic or contextual data. The perceptual attribute predictor may be configured to generate the predictions based on the extracted objective features and the extracted subjective features.

[0054] In various embodiments, the perceptual attribute predictor may include linear regression models configured to generate the predictions based on acoustic or psychoacoustic parameters computed based on the ambient soundscape data and each masker configuration.

[0055] In various embodiments, the perceptual attribute predictor may include one or more deep neural networks configured to generate the predictions based on raw audio, spectrogram representations computed based on the ambient soundscape data and each masker configuration, or a combination of the raw audio and the spectrogram representations.

[0056] In various embodiments, the data acquisition system may be configured to provide or generate the ambient soundscape data based on inputs from one or more microphones, one or more antennas, a storage medium, or any combination thereof.

[0057] In various embodiments, the plurality of masker configurations may include one or more audio tracks of recorded or synthesized sounds, one or more audio tracks of silence, or

one or more audio tracks derived from the one or more audio tracks of recorded or synthesized sounds and the one or more audio tracks of silence.

[0058] In various embodiments, the playback system may include one or more speakers, one or more virtual or augmented reality headsets, one or more headphones, one or more earbuds, or any combination thereof.

[0059] In various embodiments, the predictions generated may be non-deterministic.

[0060] FIG. 3 shows a schematic of a soundscape augmentation system according to various embodiments. The soundscape augmentation system may also be referred to as an automatic masker selection system (AMSS). The soundscape augmentation system may include a data acquisition system 302 that is responsible for providing ambient soundscape data 312, i.e. audio of an acoustic environment and/or parameters directly computed from the audio of the acoustic environment. In order to obtain the ambient soundscape data 312 from the acoustic environment, the data acquisition system 302 may use, for instance, a real-time recording from one or more microphones, a digital audio signal received from a wireless antenna, data stored on attached storage media, or any combination thereof.

[0061] Additionally, the soundscape augmentation system may include a database including a plurality of masker configurations 304 (alternatively referred to as bank of candidate masker configurations). The term “masker configuration” may refer to any audio track that could be added to the acoustic environment in 312 via any form of playback, in tandem with other parameter(s) or effect(s) necessary for intended playback of the audio track. A masker configuration may alternatively be referred to as “masker”. Such parameter(s) or effect(s) may include, for instance, digital gain levels, spatialization, dynamic envelope, and/or the application of filtering. Examples of masker configurations may include, but may not be limited to audio tracks consisting of recorded or synthesized sounds played at different volume levels, audio tracks consisting purely of silence, whose playback would constitute no addition of sound

to the acoustic environment in 312, and/or audio tracks derived from a function of one or more other audio tracks in 304. Masker configuration may alternatively or additionally refer to the process of selecting the masker, for instance from a database 304 of audio tracks. The masker configuration strategy may be autonomous and may adapt to the dynamics (e.g. amplitude, frequency content, context) of the present soundscape (e.g. via a monitoring microphone, or by digital input).

[0062] The data acquisition system 302 may optionally also be responsible for capturing demographic or contextual data 314 as auxiliary parameters to the acoustic environment itself, and for providing the demographic or contextual data 314. Such demographic or contextual data 314 may include, but may not be limited to (1) non-acoustic environmental parameters such as location (via a GPS sensor), number of people nearby (via a camera), air temperature (via a thermometer), humidity (via a hygrometer) and/or wind speed (via an anemometer); (2) demographic data of real or hypothetical listeners in the acoustic environment (e.g. via a digital form attached to the system or as pre-loaded data on a storage device), such as age, gender, and/or occupation; (3) psychological data of real or hypothetical listeners in the acoustic environment (e.g. via similar methods as demographic methods), such as noise sensitivity (e.g. using the Weinstein Noise Sensitivity Scale), perceived stress (e.g. using Cohen's Perceived Stress Scale), and/or well-being index scores (e.g. using the WHO-5 Well-Being Index); and/or physiological data of real or hypothetical listeners in the acoustic environment (e.g. via a portable health monitoring device), such as heart rate, blood pressure and/or body temperature.

[0063] The ambient soundscape data 312 and optionally the demographic or contextual data 314 may be provided by the data acquisition system 302 to the perceptual attribute predictor 306. Given the ambient soundscape data 312, masker configurations from the database 304, and optionally the demographic or contextual data 314 as inputs, the perceptual attribute predictor 306 may output predictions (of listeners to which the data in 314 corresponds to) on

one or more pre-defined perceptual attribute scales for each masker configuration in the database 304, as if its playback is to be realized in the acoustic environment. The one or more pre-defined perceptual attribute scales may, for instance, be from the circumplex octant scale from ISO12913-3 for each masker configuration in the database 304. The one or more pre-defined perceptual attribute scales may, for instance, be based on pleasantness, eventfulness, sound quality, perceived loudness, sharpness, roughness etc. for each masker configuration in the database 304. FIG. 4 shows the two-dimensional circumplex octant model of affective quality attributes as presented in ISO12913-3 in which predictions from the perceptual attribute predictor 306 according to various embodiments may be based on.

[0064] The collection of such predictions is denoted by 316 in FIG. 3. The predictions 316 may or may not be single numerical values. Similarly, in various embodiments, the predictions 316 generated may be non-deterministic, while in various other embodiments, the predictions 316 generated may be deterministic. The term “deterministic” may mean that the same output is produced whenever the same inputs from the ambient soundscape data 312, the plurality of masker configurations from the database 304 and optionally the demographic or contextual data 314 are provided to the perceptual attribute predictor 306. A probability distribution (or random values drawn from it), a vector, or any other multivariate representation of the perceptual attribute(s) may be possible outputs as predictions 316.

[0065] The perceptual attribute predictor 306 may be or may include any model taking in same inputs from the ambient soundscape data 312, the plurality of masker configurations from the database 304 and optionally the demographic or contextual data 314 to output predictions 316. Such models may transform the ambient soundscape data 312, the plurality of masker configurations from the database 304 and the demographic or contextual data 314 internally before delivering the predictions 316 as output. Examples of such models may include, but may not be limited to: (1) linear regression models taking in acoustic or psychoacoustic parameters

computed from the ambient soundscape data 312, the plurality of masker configurations from the database 304 and aggregated mean values computed from the demographic or contextual data 314 to output the predictions 316; (2) deep neural networks taking in representations computed from the ambient soundscape data 312, the plurality of masker configurations from the database 304, concatenated with representations of the demographic or contextual data 314 to output the predictions 316; (3) composite networks as shown in FIGS. 5 – 7; and/or (4) any combination as described herein.

[0066] FIG. 5 is a schematic of the perceptual attribute predictor 506 according to various embodiments. In various embodiments, the perceptual attribute predictor 506 may be configured to generate the predictions by combining the ambient soundscape data and each masker configuration, e.g. in an inference engine at pre-calibrated soundscape-to-masker ratios. A masker waveform 504a of each masker configuration may be weighted (i.e. multiplied) with a masker gain input 504b, e.g. a masker gain waveform, to generate a weighted masker waveform. The weighted masker waveform may be combined with (i.e. added to) an ambient soundscape waveform 512 of the ambient soundscape data to generate an augmented soundscape waveform 518. The augmented soundscape waveform 518 may be received by a prediction block 520 of the perceptual attribute predictor 506 to generate the predictions 510.

[0067] Various other embodiments may not require mixing, combining or adding the ambient soundscape data and each weighted masker waveform to form the augmented soundscape waveform before the augmented soundscape waveform is being inputted into the perceptual attribute predictor. By not requiring the mixing, combination or addition of the ambient soundscape waveform and each weighted masker waveform to form the augmented soundscape data before the augmented soundscape waveform is being inputted into the perceptual attribute predictor, various embodiments may allow for a more compute- and bandwidth-efficient system in real-world deployment.

[0068] FIG. 6 is a schematic of the perceptual attribute predictor 606 according to various embodiments. In various embodiments, the perceptual attribute predictor 606 may include a soundscape feature extractor 622 configured to extract features 624 of the ambient soundscape data, i.e. extract features 624 of an ambient soundscape waveform 612 of the ambient soundscape data. The perceptual attribute predictor 606 may include a masker feature extractor 626 configured to extract features 628 of each masker configuration, i.e. extract features 628 of a masker waveform 604a of each masker configuration. The perceptual attribute predictor 606 may include a feature level augmentor 630 configured to generate one or more augmented soundscape features 632 based on the extracted features 624 of the ambient soundscape data, the extracted features 628 of each masker configuration and a masker gain input 604b, e.g. a digital gain level. The perceptual attribute predictor 606 may include a prediction block 620 configured to generate the predictions 610 based on the one or more augmented soundscape features 632, the extracted features 624 of the ambient soundscape data and the extracted features 628 of each masker configuration. In summary, the perceptual attribute predictor 606 may be configured to generate the predictions 610 based on the extracted features 624 of the ambient soundscape data and the extracted features 628 of each masker configuration.

[0069] FIG. 7 is a schematic of the perceptual attribute predictor 706 according to various embodiments. In various embodiments, the perceptual attribute predictor 706 may include an objective feature block 734 configured to extract objective features 736 from the ambient soundscape data and each masker configuration, i.e. from an ambient soundscape waveform 712 of the ambient soundscape data and a masker waveform 704a of each masker configuration. The perceptual attribute predictor 706 may be further configured to extract the objective features 736 also based on a masker gain input 704b. The perceptual attribute predictor 706 may include a subjective feature block 738 configured to extract subjective features from the demographic or contextual data 714. The subjective feature block 738 may also be configured

to process the extracted subjective features and the extracted objective features 736 to generate the predictions 710. The perceptual attribute predictor 706 may be configured to generate the predictions 710 based on the extracted objective features 736 and the extracted subjective features.

[0070] Referring back to FIG. 3, the predictions 316 provided by the perceptual attribute predictor 306 may be inputted into a masker configuration ranking system 308 to generate a single optimal masker configuration 340. The optimality may be determined by the masker configuration ranking system 308 via any suitable metric, which may be but is not limited to the maximum or minimum of perceived loudness, sound quality, sharpness, roughness, one or more affective quality attributes shown in FIG. 4, and/or any combination thereof. Alternatively, it may be envisioned that the predictions 316 provided by the perceptual attribute predictor 306 may be inputted into the masker configuration ranking system 308 to generate multiple optimal masker configurations. There may be multiple optimal masker configurations for a given metric (e.g. if there are multiple masker configurations tied in the metric decided upon by the user).

[0071] The optimal masker configuration 340 may be used as input to a playback system 310. The playback system 310 may realize the actual playback or reproduction of the optimal masker configuration 340 in the acoustic environment. The playback system 310 may include, but may not be limited to, one or more speakers, one or more virtual or augmented reality headsets, one or more headphones, one or more earbuds, and/or any combination thereof.

[0072] Various embodiments may be used in outdoor urban areas with poor sound quality (e.g. a park next to a highway) to mitigate the perceived adverse impact of noise in the surrounding acoustic environment, or indoor areas (e.g. the interior of a nursing home) to improve the acoustic comfort and overall quality of the surrounding acoustic environment. Since the system effectively aims to alter the perception of a surrounding acoustic environment,

it may be of interest to any commercial entity owning an indoor or outdoor area whose acoustic environment needs to be maintained in a fixed or desired condition.

[0073] Various embodiments may be capable of choosing and adding maskers not just to real, but hypothetical ambient soundscapes. Various embodiments may also be used to create abstract "experiential zones" by adding maskers if arbitrary audio is fed into the data acquisition system instead of real-time recordings from a microphone. This would be of interest to any commercial entity interested in crafting or modifying soundscapes in virtual, mixed, or augmented reality based on listener perception.

[0074] Various embodiments may relate to a system that selects and configures an audio track (i.e. a masker) to be introduced to an existing soundscape (e.g. by playback through loudspeakers, or by digital addition via headphones) based on optimizing predicted perceptual responses (e.g., perceived loudness, circumplex octant scale from ISO 12913-3 – pleasantness, annoyance, tranquillity, calmness, vibrancy, eventfulness, etc.) to the mixed/augmented soundscape.

[0075] Various embodiments may choose masker configurations or maskers to add to the ambient acoustic environment based on predictions of human perception made by a perceptual attribute predictor, which outputs values or distributions of perceptual attribute scales given ambient soundscape data (real or hypothetical), data of candidate masker configurations, and optionally demographic or contextual data on listeners of that ambient soundscape. This embodies the ISO 12913 definition of "soundscape" to consider acoustic environments as perceived by a person or people and in context. In contrast, existing methods may choose masker configurations or maskers based on objective acoustic parameters (e.g. spectrum, sound pressure level, types and directions of sound sources) of the ambient acoustic environment and may not consider the subjective effect of human perception into masker choice or masker configuration choice.

[0076] In various embodiments, the perceptual attribute predictor may output perceptual predictions in near real time (<10s latency between input and predictions) and hence automatically suggest time-varying optimal masker configurations for a time-varying acoustic environment in real time. In other words, various embodiments may autonomously change the maskers (i.e. the audio tracks) and/or the configurations (e.g. gain levels, spatialization) being played as the acoustic environment changes over time. In contrast, existing systems only allow for static masker configurations (with only dynamic volume/filter effects potentially being applied) or user-controlled masker configurations (i.e. a person must manually change the configuration based on their own personal/expert opinion of optimality).

[0077] In various embodiments, the users may set the system to automatically pick optimal maskers on a variety of selection schemes. The selection scheme may be arbitrary, but various embodiments may allow for the following selection schemes that are absent from conventional solutions: (1) probabilistic top-k: rank the predicted perceptual response values of the combined maskers and existing soundscape across different maskers and randomly pick one among the k masker configurations giving the top (or bottom) k ranked values; (2) probabilistic distribution: draw predictions of perceptual response values of the combined maskers and existing soundscape as random samples from probability distributions generated by the system.

[0078] Existing systems and masker selection strategies may not allow optimal maskers to be chosen via a dynamic strategy, and may feature one of the following: (1) manual selection: either the user or administrator of the system chooses optimal masker based on their own opinion; (2) deterministic: pick the optimal masker as a deterministic, predefined function of the perceptual response values of the combined maskers and existing soundscape across different maskers.

[0079] Various embodiments may include one or more audio transducers (e.g. loudspeakers, headphones) and one or more sensors (e.g. microphones, environment sensors), which are

intercoupled to one or more computing devices. Various embodiments may have advantages not possessed by existing systems/solutions.

[0080] For instance, various embodiments may employ non-energetic masking. Energetic masking leverages on the physical limitations of the peripheral human auditory processes. The amplitude and frequency bandwidth of the masking sound is adjusted such that the target noise becomes completely or partially inaudible at the inner ear. Hence, objective metrics such as sound pressure level or psychoacoustic parameters are often employed to optimise or select energetic maskers. Various embodiments may take reference from (but are not limited to) the concept of soundscapes, as defined in ISO 12913-1. The augmentation of soundscapes with additional sounds is also commonly referred to as masking in the literature. The masking technique may include both energetic and informational (or perceptual) masking, wherein informational masking consists of effects concerning the noticeability of sounds that are associated with audio in the higher brain centres. Hence, the maskers may be determined by perceptual factors (e.g. perceived affective attributes in ISO 12913-2, perceived annoyance) and not just solely on objective measures as that in existing systems/solutions.

[0081] The advantage of non-energetic masking may be evidenced in the literature on masking of traffic noise with natural sounds, whereby the perceived loudness reduction of energetic masking becomes nullified as the target noise levels increase. This is opposed to a statistically significant perceived loudness reduction and soundscape quality improvement when the maskers providing informational masking are presented at up to 6 dB lower levels at high levels of traffic noise. Moreover, similar significant reduction of perceived loudness and improvement in soundscape quality was observed with birdsongs as maskers, which are unable to energetically mask traffic noise, lending further credence to the effectiveness of non-energetic masking.

[0082] The ability to contextualise the audio input via an auxiliary input may allow various embodiments to choose maskers for an arbitrary target soundscape (instead of just being limited to office environments described in existing literature). The maskers according to various embodiments may also not be limited to tracks generated from random noise nor restricted to natural sounds as described in existing literature. There may also be no restriction on the number of maskers that can be added as long as the desired perceptual attribute is optimised (which can mean "maximised" or "minimised").

[0083] In various embodiments, the audio input data may be format agnostic and may be real-time data from physical microphones or an audio track accessed from storage media. Similarly, the maskers according to various embodiments may be streamed to loudspeakers (or headphones) or digitally mixed with the target soundscape from storage media and played as a single combined track in a virtual environment. Hence, various embodiments may be able to mask soundscapes in real-world environments as well as virtual environments (e.g. in metaverse applications).

[0084] Dynamic masking described in existing literature is either based on objective metrics, such as sound level, audio frequency spectra, or by proximity to the identified target noise source. Various embodiments may dynamically or adaptively adjust masker configurations based on predicted perceptual attributes. Masker configurations may not be limited to sound levels or spectral adjustments.

[0085] Example Study 1

[0086] Introduction

[0087] It is well known that indicators based purely on the sound pressure level (SPL) of a given acoustic environment are normally insufficient in reflecting the level of annoyance and the impact to the quality of life caused by excessive noise. This has led to the rise of the soundscape approach to noise control, which focuses on interventions that improve perceptual

attributes of noise instead of simply reducing the SPL. The ISO 12913 series of international standards on soundscapes aims to codify this approach by providing a circumplex model of pleasantness and eventfulness, upon which interventions can be compared, based on subjective evaluations of the surrounding acoustic environment.

[0088] Consequently, many studies have utilized soundscape augmentation techniques to alter a perception of a soundscape by adding maskers to an urban or indoor acoustic environment to optimize metrics, such as its subjectively-evaluated pleasantness or calmness. However, the choice of maskers is usually arbitrary, expert-guided, or based on post-hoc analysis. An arbitrary choice of masker may be unreliable in effecting a desired perceptual change, an expert-guided choice is labour- and time-intensive, and post-hoc analyses may not be generalizable to unseen maskers and soundscapes in an unobserved context.

[0089] One way to overcome these limitations is to train a prediction model on an acoustically diverse selection of soundscapes and maskers to predict the value of some perceptual attribute, as subjectively evaluated by a person, given the raw auditory stimuli. Once trained, simulated additions of maskers to an unseen soundscape can be fed as input to the model to obtain predictions of said perceptual attribute. Then, the masker effecting the greatest increase or decrease in the attribute can be selected for in-situ augmentation as the optimal masker.

[0090] Various embodiments may relate to a neural approach for an automatic masker selection system (AMSS) by optimizing the pleasantness of an acoustic environment to augment urban soundscapes. Using a probabilistic output scheme, the models predict a distribution for the pleasantness for each soundscape-masker combination, rather than a single deterministic value, allowing for the confidence level of a prediction to be explicitly retrieved along with the predicted pleasantness.

[0091] Related Work

[0092] Studies developing prediction models for perceptual attributes in soundscape research have primarily focused on simpler machine learning models, such as linear regression, support vector machines (SVM), and shallow multilayer perceptrons. They use input features based on acoustic measurements, psychoacoustic parameters, environmental characteristics, or some linear combination of them elucidated by principal component analysis. A scaling metric for Likert scales to account for nonlinearities in the scales used to develop these systems has also been proposed.

[0093] On the other hand, a recent systematic review showed that studies making use of deep neural networks to predict perceptual attribute values of soundscapes are rare, despite their prevalence in more “objective” tasks, such as sound event localization, detection, and classification. The most significant study in this respect appears to be one comparing the performances of SVM, a convolutional neural network (CNN), long short-term memory network (LSTM), and a fine-tuned VGGish network on the Emo-Soundscapes dataset. Emo-Soundscapes contains 1213 audio clips ranked by valence (pleasantness) and arousal (eventfulness), according to subjective paired comparisons via the Self-Assessment Manikin. The fine-tuned VGGish model and the CNN trained from scratch were respectively found to have the least mean squared errors (MSE) in predicting the valence and arousal of the audio clips. A more recent study made use of similar models to classify music into one of four quadrants in a circumplex model similar to that in ISO 12913 and into the categories neutral, calm, happy, sad, angry, and fearful, using datasets totalling about a thousand audio clips. Soundscapes, however, usually contain more than just music, so it remains to be investigated if the results are generalizable to broader categories of acoustic stimuli.

[0094] In addition, these existing models in the literature are deterministic models, in the sense that the same acoustic environment is always mapped to the same predicted value without accounting for the inherent uncertainties in the subjective ratings used as ground truths.

However, these ratings given by individual people are inherently random due to factors that cannot be reasonably controlled, such as the person's current stress level or noise sensitivity. As such, various embodiments may relate to a probabilistic approach by training a neural network to predict a distribution across possible ratings of the output and subsequently drawing from it, allowing the model to account for the varying levels of uncertainty in each unique soundscape.

[0095] Proposed Method

[0096] FIG. 8 shows a training and inference schema of an automatic masker selection system (AMSS) according to various embodiments.

[0097] Consider a soundscape S_0 and a set of K candidate maskers $\{M_k\}$ that can be added to the soundscape. Denote by S_k the augmented soundscape with masker M_k . In other words, S_k is the soundscape S_0 after the addition of M_k , accounting for the response of the play-back system and the real-world environment. Finding the optimal $\tilde{M}$ to maximize the evaluation of a given perceptual (or affective attribute) f on the augmented soundscape is equivalent to finding $\operatorname{argmax}_k f(S_k)$, where f could, for instance, be the rating of pleasantness, eventfulness, or some other perceptual attribute on a numerical scale. We can then augment S_0 with $\tilde{M}$ to obtain $\tilde{S}$.

[0098] A naive implementation of an AMSS could, for instance, use some approximator of $f(\cdot)$ to output a predicted value $\hat{f}_k := \hat{f}(S_k)$ of the perceptual attribute for each augmented soundscape S_k and pick the masker with the highest $\hat{f}_k$. However, as mentioned earlier, this deterministic output scheme neglects the inherent uncertainty associated with subjective ratings on perceptual attribute scales. One soundscape, for instance, could be very pleasant to some but very annoying to others, while another soundscape could be nearly universally mildly pleasant to any listener, depending on the context.

[0099] To capture this uncertainty, f may be treated as a random variable with a joint distribution $p(f, S)$ with some “random” soundscape S , where the observed augmented soundscape S_k is treated as a realization of S . The uncertainty would then be represented in the variance of f . A prediction model in this scenario would then be attempting to model the conditional distribution $p(f|S = S_k)$ as some function of $\hat{F}(S_k)$ of the augmented soundscape S_k . $\hat{F}(S_k)$ is shortened to $\widehat{F}_k$ for brevity.

[0100] A neural approximator 806 may be used to output $\widehat{F}_k$. The neural approximator 806 may be termed as a probabilistic perceptual attribute predictor (PPAP). In the proposed AMSS shown in FIG. 8, each masker may be passed to the aforementioned PPAP 806, which outputs one or more parameters of some predefined distribution family. A PPAP 806 may be trained to output the mean μ_k and the log standard deviation $\log \sigma_k$ of a normal distribution $N(\mu_k, \sigma_k^2)$, which is used to model the output $\widehat{F}_k$. Once sufficient (or all) predicted distributions $\widehat{F}_k$ are computed, the “best” masker may be selected based on some pre-decided criteria. For example, a masker may be selected simply based on the highest μ_k , a set of deterministic values akin to $\{\widehat{f}_k\}$ could be sampled from $\{\widehat{F}_k\}$ to encourage masker exploration, or a more sophisticated criterion taking σ_k into account could be used.

[0101] Although only the observed values of f are available and the true distribution of f is unknown, the model still can be optimized by maximizing the log-probability of the ground truth given the output distribution, in a manner inspired by Bayesian optimization. Given a soundscape S_0 , a set of maskers $\{M_k\}$, and ground truths $\{f(S_k)\}$, the contribution to the loss function of the soundscape may be given by

$$\mathcal{J}_{prob.}(\{M_k\}; \{f(S_k)\}) = -\frac{1}{K} \sum_k \mathcal{L}(f(S_k); \mu_k, \sigma_k) \quad (1)$$

$$= \frac{1}{K} \sum_k \left[\frac{1}{2} \left(\frac{f(S_k) - \mu_k}{\sigma_k} \right)^2 + \log \sigma_k \right] \quad , \quad (2)$$

where $\mathcal{L}$ is the log density function of the output distribution $N(\mu_k, \sigma_k^2)$, omitting additive constants. During training, the model may be optimized through batches of soundscape-masker pairs with available ground truths.

[0102] As seen in Equation (2), the loss function using the normal distribution can be considered as a weighted MSE loss regularized by the log standard deviation. This loss function is inherently stable with respect to σ_k as the first term encourages larger σ_k , while the second term encourages smaller σ_k . Of course, some other choices of the output distribution may reduce similarly to other deviation measures, such as the Laplace distribution reducing to a regularized mean absolute error. Note that it is also possible to force to the model to learn deterministically by setting σ_k to some predetermined constant and training the model using the pure MSE loss between the ground and μ_k . For the purpose of an ablation study, the following may be used

$$\mathcal{J}_{det}(\{M_k\}; \{f(S_k)\}) = \frac{1}{2K} \sum_k (f(S_k) - \mu_k)^2 \quad (3)$$

as the deterministic counterpart to Equation (2). The deterministic loss in Equation (3) may be thought of as Equation (2) with a static $\sigma_k = 1$.

[0103] Validation Experiments

[0104] To validate the proposed system, we let f be the normalized pleasantness measure as defined by the ISO 12913-3 standard, which may be referred to as ISO Pleasantness. Specifically,

$$f(S_k) = \frac{2(r_{pl}-r_{an})+\sqrt{2}(r_{ca}-r_{ch}+r_{vi}-r_{mo})}{8(1+\sqrt{2})} \in [-1,1] \quad (4)$$

where $r_{pl}, r_{an}, r_{ca}, r_{ch}, r_{vi}, r_{mo} \in \{1,2,3,4,5\}$ are the extent to which a participant considers the augmented soundscape S_k to be pleasant, annoying, calm, chaotic, vibrant, and monotonous, respectively, on a 5-point Likert scale. For each S_k , the distribution $\widehat{F}_k = N(\mu_k, \sigma_k^2)$ may be predicted, with the ‘ground-truth’ labels $f(S_k)$ being the ISO Pleasantness rating for S_k given by the participant. As an ablation study, models trained with the proposed method against

deterministic models that predict $f(S_k)$ via μ_k directly are compared. The validation experiments were conducted on a dataset of subjective responses to a variety of augmented soundscapes.

[0105] Dataset

[0106] The dataset contains 12 600 unique sets of subjective responses to augmented soundscapes in a 5-fold cross-validation set, as well as 48 additional sets of responses in an independent test set, for a total of 12 648 samples. Each sample maps an augmented soundscape (as a raw audio recording) to an ISO Pleasantness value given the set of responses to the following 6-item subset of the ISO 12913-2 affective response questionnaire: “To what extent do you agree or disagree that the present surrounding sound environment is {pleasant, chaotic, vibrant, calm, annoying, monotonous}?”

[0107] Participants responded on a 5-point scale with the labels “Strongly disagree”, “Disagree”, “Neither agree nor disagree”, “Agree”, and “Strongly agree”, which were respectively coded as 1, 2, 3, 4, and 5. The coded responses were then used in Equation (4) to compute the ground truth labels of ISO Pleasantness for each augmented soundscape.

[0108] Augmented Soundscapes

[0109] The augmented soundscapes in the 5-fold cross-validation set were made by adding 30-second excerpts of recordings from Freesound and xeno-canto as “maskers” to 30-second excerpts of binaural recordings of the soundscapes from the Urban Soundscapes of the World (USotW) database, a comprehensive dataset of urban soundscapes. The unaugmented soundscapes were also included as controls. All recordings used were sampled at 44.1 kHz.

[0110] Each fold has a bank of 56 maskers in the following classes: bird (16), construction (8), traffic (8), water (16), and wind (8). The classes were chosen to cover the range of sound types generally evaluated to be pleasant and annoying. Maskers and soundscapes in each fold are disjoint.

[0111] The augmented soundscapes in the test set were made in a similar fashion with 7 maskers (8 including the unaugmented control) independent of those from the cross-validation set and 6 binaural recordings of soundscapes independent of the USotW dataset, recorded using the same Soundscape Indices Protocol. The 7 maskers were excerpted from xeno-canto track IDs (identification numbers) 640568 (bird “B1”) and 568124 (bird “B2”), as well as Freesound track IDs 586168 (construction “Co”), 587219 (traffic “Tr”), 587000 (water “W1”), 587759 (water “W2”), and 587205 (wind “Wi”). The maskers were added exhaustively to all soundscapes for the test set, resulting in 48 samples in the test set. All soundscapes were calibrated according to the method described in *K. Ooi et al.* (“Automation of binaural headphone audio calibration on an artificial head,” *MethodsX*, vol. 8, no. February, p. 101288, 2021) to the measured in-situ A-weighted equivalent SPL ($L_{A,eq}$) before adding the maskers at a constant soundscape-to-masker ratio of 0 dB for the test set, and a randomly-selected value, in dB, from $\{-6, -3, 0, 3, 6\}$ for the cross-validation set.

[0112] Subjective Responses

[0113] To obtain subjective responses to the augmented soundscapes, 300 participants were recruited to each rate 42 unique, randomly-chosen augmented soundscapes from a fold of the validation set. 5 other participants were recruited to each rate the 48 augmented soundscapes in the test set. Each augmented soundscape in the validation and test set is therefore rated by one and five participants, respectively. All participants listened to the calibrated augmented soundscapes on a pair of circumaural headphones (Beyerdynamic Custom One Pro), powered by an external sound card (Creative SoundBlaster E5). After listening to each augmented soundscape, they answered the 6-item questionnaire described earlier (under the header “Dataset”), and the ISO Pleasantness was calculated from their responses according to Equation 4.

[0114] Model And Training

[0115] Log-mel spectrograms of the augmented soundscapes were used as inputs. The log-mel spectrograms were extracted using a 4096-sample Hann window with 50 % overlap and compressed to 64 mel bins. FIG. 9 shows (a) the base convolutional recurrent neural network (CRNN) architecture used according to various embodiments; (b) one possible implementation of the feature mapping block in (a) according to various embodiments; and (c) other possible implementations of the feature mapping block in (a) according to various other embodiments. Four different feature mapping blocks, namely a vanilla mapping block using a bidirectional gated recurrent unit (BiGRU), an additive attention block, a dot-product attention block, and a multi-head attention block with 4 heads were investigated. The vanilla mapping block is shown in FIG. 9(b), while all attention-based blocks sharing the same general workflow are shown in FIG. 9(c). The last layers of the model in FIG. 9(a) are dense layers which finally output μ_k and $\log \sigma_k$. In the deterministic ablation models, $\log \sigma_k$ is ignored.

[0116] All models were trained using a 5-fold cross-validation scheme, with 10 models per fold, to a total of 50 models. Each fold uses the same 10 seeds for the models. All models were trained up to 100 epochs using an Adam optimizer with a learning rate of 5×10^{-5} . For each model, the model weights with the best validation loss are used for evaluation in both the cross-validation set and the test set.

[0117] Results And Discussion

[0118] FIG. 10 is a table showing the mean fold mean squared errors (MSEs) of the probabilistic perceptual attribute predictor (PPAP) ($\pm$ standard deviation) over the 10 runs tested for each setting (above: cross-validation set, below: test set) according to various embodiments. Each run may include five models, each on a different fold of the cross-validation set but with the same initial conditions. For all models, the contributions to the MSEs were calculated as $(\mu_k - f_k)^2$. Asterisks (*) denote statistically significant improvements ($p < 0.05$). FIG. 10 summarizes the results of the validation experiments described earlier on both

the 5-fold cross validation set and the independent test set. As a reference, the results of a trivial, deterministic “label mean model” are also provided. In this “label mean model”, the mean of labels in the training set is used as the prediction for all stimuli in the validation and test set. All other investigated models performed better than the label mean model, thus indicating meaningful feature extraction by the models. This is likely due to the fact that, compared to the validation set (12 600 samples, 280 maskers), the test set was smaller and less diverse (48 samples, 7 maskers).

[0119] Furthermore, for the four architectures tested, the models with probabilistic output performed better than the identical models with deterministic output for both the validation and test set. The percentage reduction in MSE for each model is shown in FIG. 10 as well. It may be observed that the CRNN with the multi-head attention block and the vanilla CRNN respectively experienced the greatest improvement in validation and test set MSE of 1.1 % and 7.8 %.

[0120] To quantify the significance of these reductions, two-sided Wilcoxon signed-rank tests between the deterministic models and probabilistic models were performed for both the validation set MSEs and test set MSEs. The reductions that were significant at a 5 % significance level are marked with asterisks in FIG. 10. The reductions in the test set MSEs for the vanilla CRNN and PPAP with additive attention are statistically significant and provide evidence to support the generalizability of the proposed method to unseen data. This may be because the inherent randomness contributed by the test set participants differs from participants in the training set, and the PPAP generalizes over this randomness better. Nonetheless, a further ablation study would be required to validate this hypothesis.

[0121] However, the lowest improvement in test set MSE was actually observed in the PPAP with multi-head attention block. This may be due to the fact that the PPAP with multi-head attention block has about 50 % more parameters than the other three models (about 120K

against about 80K), which could have caused the trained models (both deterministic and probabilistic) to be overfitted to the comparatively smaller dataset of 12 600 training and validation samples.

[0122] FIG. 11 illustrates masker selection using the 50 models (dot-product variant) on the test set, with a naïve maximum μ_k selection scheme (left) according to various embodiments and a random sampling scheme to encourage masker exploration (right) according to various embodiments. The selection criterion for each base soundscape is shown on top of each subplot. The cell colour indicates the average ISO Pleasantness rated by the test-set participants. The circle size indicates the number of models selecting the masker. The masker descriptions have been provided earlier (under the header “Augmented Soundscapes”).

[0123] In the naïve selection scheme, most models select the first bird masker (B1) across all base soundscapes. This is consistent with the consensus in previous soundscape studies, which generally observed more pleasant responses to bird-song maskers. However, the second bird masker (B2) might have generally received lower ratings by participants and correspondingly low vote counts by the models due to it being considered less pleasant in the context of the test set soundscapes, which has been observed in a previous study.

[0124] In the random selection scheme, the system still tends to select B1 which gives a good pleasantness score for most of the soundscapes, but is now exploring other maskers significantly more than the naïve scheme. This can be useful in a real-time system where human feedback can be obtained in-situ to adaptively adjust the masker selection or improve future models. It can also be seen that with Soundscape 5, where the Pleasantness of the unaugmented soundscape is already relatively high, the exploration rate is higher than other base soundscapes, allowing a more diverse acoustic experience with only a small compromise on the pleasantness level.

[0125] Conclusion

[0126] Various embodiments may relate to an automatic masker selection system (AMSS) for human-centric urban soundscape augmentation, using a probabilistic perceptual attribute predictor (PPAP). The proposed PPAP was implemented using a convolutional recurrent neural network and trained to output predictions in a probabilistic manner. This allowed it to simultaneously predict the perceptual attribute of a soundscape while accounting for the inherent randomness in human subjective perception of acoustic stimuli. Via a large-scale listening test with more than 300 participants and more than 12K unique soundscapes, we validated the effectiveness of our PPAP in predicting the pleasantness of augmented soundscapes, including those generated from unseen soundscapes and maskers. Future works on the AMSS may include in-situ implementations to assess its ecological validity, as well as investigation of the proposed method on other perceptual attributes (f), such as eventfulness or calmness, because the proposed method is not specific to any particular attribute. Indeed, since the primary assumption underpinning the PPAP is that of random ground-truth labels, one may also conceivably apply it to any context where predictions of subjective evaluations are desired.

[0127] Soundscape augmentation, which involves the addition of sounds known as “maskers” to a given soundscape, is a human-centric urban noise mitigation measure aimed at improving the overall soundscape quality. However, the choice of maskers is often predicated on laborious processes and is inflexible to the time-varying nature of real-world soundscapes. Owing to the perceptual uniqueness of each soundscape and the inherent subjectiveness of human perception, a probabilistic perceptual attribute predictor (PPAP) has been proposed that predicts parameters of random distributions as outputs instead of a single deterministic value. Using the PPAP, an automatic masker selection system (AMSS), which selects optimal masker candidates based on the predicted distribution of the ISO 12913-3 Pleasantness score for a given soundscape has been developed. Via a large-scale listening test with 300 participants, 12600

subjective responses have been collected, each to a unique augmented soundscape, to train the PPAP models in a 5-fold cross-validation scheme. Using a convolutional recurrent neural network backbone and experimenting with several variants of the attention mechanism for the PPAP, evaluation of the proposed system has been carried out using a blind test set with 48 unseen augmented soundscapes to assess the effectiveness of the probabilistic output scheme over traditional deterministic systems.

[0128] Example Study 2

[0129] Introduction

[0130] Mitigation of urban noise pollution is a complex and multifaceted problem where elimination or reduction of noise sources is often not a practical option. Unlike the traditional noise control approach aimed at reducing the sound pressure level or acoustic energy, the soundscape approach presents a more holistic and human-centric strategy based on perceptual acoustic comfort. One such technique commonly termed soundscape augmentation involves the addition of “wanted” sound(s) into the acoustic environment in order to “mask” the noise and improve the overall perceptual acoustic quality, with promising results in both virtual and real settings across various types of noise sources.

[0131] However, the selection of “maskers” and the playback levels has traditionally been either arbitrary, expert-guided, or based on post-hoc analysis. Not only are these masker selection processes often time-consuming and labour-intensive, but they are also inflexible to the dynamic nature of real-world soundscapes. As a result, it is often unlikely that these static choices of maskers and playback levels are able to consistently achieve optimal acoustic comfort. To address these limitations, Example Study 1 presented one of the first attempts at adaptive soundscape augmentation, through the use of a masker selection system powered by a probabilistic deep learning model that predicts a probability distribution of “pleasantness” for a particular augmented soundscape.

[0132] Although the above described probabilistic perceptual attribute predictor (PPAP) model presents a convincing proof of concept for automatic soundscape augmentation powered by deep learning, the model may not be made for in-situ deployment by design. The implementation of this PPAP model requires a premixed track of augmented soundscape as an input. The mixing process was done at pre- calibrated soundscape-to-masker ratios relative to known in-situ SPL levels of the recorded unaugmented soundscape tracks. This may result in a model that is blind to masker gain level — a crucial playback parameter for in-situ deployment. Moreover, requiring a mixture of the base soundscape recording with a masker as an input adds unnecessary complexity to the overall system. First, all candidate maskers would require a lookup table of the digital gain required for a playback at a specific soundscape-to-masker ratio (SMR). Second, during the selection process, each candidate masker-SMR pair would require computing an augmented soundscape track, adding computational overhead to the system.

[0133] Various embodiments may present a PPAP model with a modified formulation which will allow for a more compute- and bandwidth-efficient system in real-world deployment. The proposed model may decouple the base soundscape, the masker, and the gain level from one another at the input stage, and instead introduce separate feature extractor branches for the base soundscape and the masker. The “augmentation” process only occurs in the feature space instead of the waveform space by using the digital gain level of the masker as a conditioning input, which is independent from the in-situ sound pressure level (SPL). An attention mechanism is then used to “query” the distribution of the target perceptual attribute based on the base soundscape feature, the masker feature, and the gain-conditioned feature. The proposed method results in complexity reduction in multiple components of the system. First, the raw waveform of the soundscape may not be required to be sent to the inference engine to be mixed with the maskers, as the less bandwidth-consuming spectral data can be used as model inputs. For cloud inference, this means significant reduction in data egress rate from the edge. Second,

the masker features can be precomputed independently from the gain level, reducing the inference time, and thus overall latency. Third, attribute prediction on multiple gain levels of the same masker can now be more quickly performed without recomputing the augmented soundscape, or even the soundscape and masker features. Lastly, addition of new maskers to the system can now be done without any need for the creation of a lookup table.

[0134] Proposed Method

[0135] Preliminaries

[0136] Consider a digital soundscape recording $\mathbf{s}_i[n] \in [-1,1]^C$, where $C \geq 1$ is the number of channels and n is the sample index. Denote the in-situ A-weighted equivalent sound pressure level, $L_{A,eq}$, in dBA, of $\mathbf{s}_i$ by $a_i = \mathcal{A}(\mathbf{s}_i)$. For brevity, all mentions of sound pressure level (SPL) will refer to the 30-second $L_{A,eq}$ unless otherwise stated. Suppose all soundscapes are recorded by the same in-situ recording setup, and assuming approximate linearity, the sound pressure-to-digital level ratio (SPDR), which is the ratio of the relative sound pressure level, in the linear scale, to the digital full scale (DFS) of the setup, can be considered roughly constant and denoted by d_0 , whose unit is $Pa/(p_0 \cdot DFS)$, where $p_0 = 20 \mu Pa$ is the reference sound pressure level. Since all soundscape inputs originate from the same recording device, this also means that the in-situ SPL information is implicitly embedded in the input spectrogram data.

[0137] Consider a single-channel masker $m_j[n] \in [-1,1]$. For the maskers, it is not assumed that they are all recorded by the same recording setup. In practice, this is due to the maskers being typically sourced from different recording systems from the in-situ recording setup, and often via open-content providers such as Freesound and Xeno-canto. As such, the SPDR of each masker is often unknown and cannot be assumed to be the same as any other. Denote the SPDR of masker m_j by d_j . Under the approximate linearity assumption (verified by exhaustive gain calibration on all maskers used in this work), a masker m_j may be normalized to the same SPDR as the soundscape recording by scaling it by d_0/d_j . If both d_j and the in-situ

SPL a_i are both known, soundscape augmentation with a particular soundscape-to-masker ratio (SMR), in dBA, can proceed relatively easily without further calibration. However, d_j may often be unknown in practice, thus a common workaround is to perform calibration to obtain the digital gain $g_j(l)$ needed to reproduce m_j at a sound pressure level of l , across multiple values of l , using a playback system with a digital level-to-sound pressure ratio (DSPR) of d_0^{-1} . The calibration required to obtain accurate values of $g_j(\cdot)$ typically may require specialized equipment, such as a recording set up with an artificial head, and can be time-consuming and/or labour-intensive, depending on the calibration method used. In this work, g_j is used explicitly as a conditioning input thus eliminating the need for the SPL-dependent SMR or the need for masker calibration.

[0138] Model

[0139] FIG. 12A shows a schematic of the probabilistic perceptual attribute predictor (PPAP) according to various embodiments. Let $\mathbf{S}_i \in \mathbb{R}^{T \times F \times C}$ be the log-mel spectrogram of the soundscape $\mathbf{s}_i$ and $\mathbf{M}_j \in \mathbb{R}^{T \times F \times 1}$ be the log-mel spectrogram of the masker m_j , where T is the number of spectrogram time frames, and F is the number of mel bins. In this work, $C = 2$, $F = 64$, and $T = 644$, corresponding of 30 s of signal sampled at 44.1 kHz with a short-time Fourier transform (STFT) window size of 4096 samples and 50% frame overlap. The soundscape and masker embeddings are respectively calculated by the feature extractor 1222 f_s and the feature extractor 1226 f_m such that

$$\mathbf{k}_i = f_s(\mathbf{S}_i) \in \mathbb{R}^{N \times D}, \quad (5)$$

$$\mathbf{q}_j = f_m(\mathbf{M}_j) \in \mathbb{R}^{N \times D} \quad (6)$$

where $D = 128$ is the embedding dimension, and $N = 20$ is the number of feature time frames. Each feature extractor 1222, 1226 may include 5 convolutional blocks, each containing a convolutional layer with 3×3 kernel and stride 1; batch normalization; dropout with

probability 0.2; Swish activation; and a 2×2 average pooling layer. The convolutional layers contain 16, 32, 48, 64, and 64 output channels in this order.

[0140] Using the query-key-value (QKV) model of attention, the perceptual attribute prediction process can be viewed as using a masker to query a mapping with the keys given by the unaugmented soundscape and the values given by the augmented soundscape. Instead of performing soundscape augmentation in the audio or spectral domain, we consider performing the “augmentation” in the embedding domain (in feature augmentation block 1230). This can be done by using a simple mapping, conditioned on the masker digital gain, such that

$$\mathbf{v}_{i,j,g} = f_g(\mathbf{k}_i, \mathbf{q}_j, \gamma) \in \mathbb{R}^{N \times D}, \quad (7)$$

where g is the querying gain level, $\gamma = \log_{10} g$, and f_g is a gain-conditioned augmentation layer. Three different implementations for f_g , namely

$$f_g^{(CAT)}(\mathbf{k}_i, \mathbf{q}_j, \gamma) = \text{Dense}^2(\mathbf{k}_i | \mathbf{q}_j | \gamma I^{N \times 1}) \quad (8)$$

$$f_g^{(ADD)}(\mathbf{k}_i, \mathbf{q}_j, \gamma) = \text{Dense}^2(\mathbf{k}_i + \gamma \mathbf{q}_j) \quad (9)$$

$$f_g^{(CONV)}(\mathbf{k}_i, \mathbf{q}_j, \gamma) = \text{Dense}(\text{Conv}(\mathbf{k}_i || \mathbf{q}_j || \gamma I^{N \times D})) \quad (10)$$

where tensor $|$ is tensor concatenation on the last axis, $||$ is the tensor stacking operator creating an additional axis, $\text{Dense}: \mathbb{R}^{N \times \bullet} \mapsto \mathbb{R}^{N \times D}$ is a dense layer with D output units, $\text{Conv}: \mathbb{R}^{N \times D \times 3} \mapsto \mathbb{R}^{N \times D}$ is a convolutional layer with D filters with the one-dimensional kernels compressing the stacked dimensions into a singleton axis. FIG. 12B illustrates three feature augmentation methods according to Equations (8), (9) and (10) in various embodiments. During training, if the masker m_j is a silent track (i.e., no augmentation), we randomize $\gamma \sim N(v, \zeta^2)$, where v and ζ are the mean and standard deviation of the log-gains of the training samples with non-silence maskers. This is to teach the model to ignore gain values for silent maskers. With the query, key, and value embeddings in place, the output embedding can be computed using any QKV-compatible attention block 1220a f_a , such that

$$\mathbf{z}_{i,j,g} = f_a(\mathbf{q}_j, \mathbf{k}_i, \mathbf{v}_{i,j,g}) \in \mathbb{R}^D. \quad (11)$$

[0141] The predicted distribution $\hat{Y}_{i,j,g}$ is then computed by the output block 1220b f_o which predicts its parameters $\mu_{i,j,g}$ and $\log \sigma_{i,j,g}$. Three types of attention may be considered, namely additive attention (AA), dot-product attention (DPA), and multi-head attention (MHA). In this example, 4 heads for MHA are used.

[0142] Loss Function

[0143] As with Example Study 1, human subjective responses may be considered to be inherently random, both in the sense that a participant is a sample of the population, and in the sense that the response given by each participant during the listening test is also a random sample of their inherently non-deterministic perception. As such, the label $y_{i,j,g}$ for the target attribute (i.e. ISO Pleasantness) may be considered an observation of a random variable $Y_{i,j,g}$ representing the distribution of the target attribute, whose true distribution is unknown. Using the maximum likelihood formulation, the optimization objective of the PPAP model can be given by:

$$\operatorname{argmax}_{\theta} \sum_{i,j,g} \mathcal{L}_{\hat{Y}_{i,j,g}}(y_{i,j,g}) \quad (12)$$

where θ is the parameters of the model and $\mathcal{L}_X(x)$ is the log-density of X evaluated at x . Modelling the subjective response to each acoustic scene by an independent normal distribution, the optimization objective in Equation (12) translates to the loss function

$$\mathcal{T} = \frac{1}{|\mathfrak{T}|} \sum_{i,j,g} \frac{1}{2} \left(\frac{y_{i,j,g} - \mu_{i,j,g}}{\sigma_{i,j,g}} \right)^2 + \log \sigma_{i,j,g}, \quad (13)$$

where $\mathfrak{T}$ is the training batch.

[0144] Optimized Inference

[0145] Although the model sees batches of different soundscape-masker-gain samples during training, during in-situ inference, the model sees only one base soundscape at time, while going through multiple masker and gain levels to find the most suitable masker-gain pair for the particular soundscape. As a result, significant performance improvement can be made by

eliminating duplicate computation instead of making a new forward pass for each masker-gain combination. Denote the time required to perform f_s, f_m and $f_g \circ f_a \circ f_o$ by τ_s, τ_m , and τ_{g+a+o} , respectively. Let η_m be the number of maskers to query per soundscape and η_g be the number of gain values to query per masker. A naive batching scheme would require a total runtime of $\eta_g \eta_m (\tau_s + \tau_m + \tau_{g+a+o})$. Instead, by optimizing the query process, as shown in Algorithm 1 of FIG. 13, the total runtime can be reduced to $\tau_s + \eta_m \tau_m + \eta_g \eta_m \tau_{g+a+o}$. FIG. 13 shows an algorithm to reduce total runtime according to various embodiments. Additionally, for a fixed masker bank, it is possible to precompute the masker features $\mathbf{q}_j$, allowing the total runtime to be further reduced to $\tau_s + \eta_g \eta_m \tau_{g+a+o}$.

[0146] Dataset

[0147] The dataset used in this work is extended with additional participants from that used in Example Study 1, totalling 18 564 data points from 442 participants (42 stimuli per participant). Each data point represents the subjective responses of a participant to an augmented soundscape.

[0148] *A. Stimuli*

[0149] The augmented soundscapes were made in 5 disjoint folds by adding 30-second “maskers” to 30-second binaural recordings of the soundscapes from the Urban Soundscapes of the World (USotW) database. The unaugmented soundscapes (i.e., soundscape “augmented” with silent maskers) were also included as controls. All recordings used were sampled at 44.1 kHz. Further details can be found in Example Study 1.

[0150] Addition of the maskers to the soundscapes were made at specific soundscape-to-masker ratios (SMR) of $\{-6, -3, 0, 3, 6\}$ dBA, where SMR is defined by the ratio of the SPL of the soundscape to that of the masker. Although the in-situ SPL of the base soundscape tracks are known, the masker data were sourced from online repositories, specifically Freesound and Xeno-canto, thus they required calibration. A lookup table of the digital gain $g_j(l)$ required to

amplify the j th masker to a specific SPL was created by exhaustively calibrating the track on a dummy head (GRAS 45BB KEMAR Head & Torso), to SPL values between 46 dBA and 83 dBA inclusive, in 1 dBA steps. The calibration process was semi-automated using the same audio playback devices as those of the participants. Since the in-situ SPL of the base soundscapes are usually non-integral, the digital gain for non-integral l is interpolated from the lookup table using

$$g_j(l) = \text{lookup}_j[\text{round}(l)].10^{(l-\text{round}(l))/20} \quad (14)$$

where $\text{lookup}_j[\lambda]$ is the calibrated gain for masker j at an SPL of λ dBA.

[0151] *B. Subjective Response*

[0152] All participants listened to the calibrated augmented soundscapes on a pair of circumaural headphones (Beyerdynamic Custom One Pro), powered by an external sound card (Creative SoundBlaster E5). After listening to each augmented soundscape, participants responded to a 6-item questionnaire based on the ISO 12913-2 standard: “To what extent do you agree or disagree that the present surrounding sound environment is [...]?”, where [...] is one of pleasant, chaotic, vibrant, calm, annoying, and monotonous. Each questionnaire item has the options “Strongly disagree”, “Disagree”, “Neither agree nor disagree”, “Agree”, and “Strongly agree”, numerically encoded from 1 to 5. The ground-truth labels of ISO Pleasantness, for each augmented soundscape is then calculated by

$$y_{i,j,g} = \frac{2(v_{pt}-v_{an})+\sqrt{2}(v_{ca}-v_{ch}+v_{vi}-v_{mo})}{8(1+\sqrt{2})} \in [-1,1] \quad (15)$$

where $v_{pt}, v_{an}, v_{ca}, v_{ch}, v_{vi}, v_{mo}$ are the numerical ratings for the pleasant, annoying, calm, chaotic, vibrant, and monotonous questionnaire item, respectively.

[0153] Experiments

[0154] For all model types in this paper, each model is trained in a 5-fold cross-validation manner for the same 10 seeds per validation fold, totalling 50 models per model type. Each model is trained for up to 100 epochs using an Adam optimizer with a learning rate of 5×10^{-5} .

The learning rate is halved if the validation mean squared error (MSE) does not decrease for at least 5 epochs and the training is stopped early if the validation MSE does not decrease for at least 10 epochs. The MSE, as well as the mean absolute error (MAE), reported herein are calculated using the difference between the mean of the predicted distribution, $\mu_{i,j,g}$, and the ground truth label $y_{i,j,g}$.

[0155] In addition to the three types of attentions mentioned above (under the header "Model"), a "pass-through" attention block which acts as a baseline model is included in the experiment. The baseline model is denoted by 'X' in FIG. 14. The pass-through block may be defined by

$$\text{PassThroughAttention}(\mathbf{q}_j, \mathbf{k}_i, \mathbf{v}_{i,j,g}) = \mathbf{v}_{i,j,g} \quad (16)$$

[0156] FIG. 14 shows (above) a plot of mean squared error (MSE) as a function of attention block type showing violin plots of the validation MSE of the pleasantness prediction for each attention block type and feature augmentation method according to various embodiments; and (below) a plot of mean absolute error (MAE) as a function of attention block type showing violin plots of the validation MAE of the pleasantness prediction for each attention block type and feature augmentation method according to various embodiments. FIG. 14 shows the distribution of the prediction errors in terms of MSE and MAE for each attention and augmentation type. Across all attention types, the CONV augmentation method performed the best in terms of both MSE and MAE, followed by CAT and ADD. This may be attributed to the CONV method naturally encouraging feature alignment between $\mathbf{k}_i$ and $\mathbf{q}_j$, and the kernel filtering allowing a more flexible augmentation operation in the feature space compared to the other two methods.

[0157] It can also be seen that using a QKV attention generally allows better performance compared to the no-attention model. However, it appears that the type of attention mechanism does not have a significant effect on the model performance, as long as some form of attention

is used to couple the query, key, and value tensors. Despite AA and DPA only having a single attention head, the four-head MHA does not seem to result in a noticeable improvement in terms of prediction error. It should also be noted that the choice of feature augmentation block seems to have a more significant impact on the accuracy of the model, as seen by the CONV augmentation models without attention generally performing better than the ADD models with attention.

[0158] Despite the model never seeing the augmented soundscape audio data presented to the participants, the best average MSE model (Luong attention with CONV augmentation) at 0.122 ± 0.005 is on par with the result in Example Study 1, which used the log-mel spectrogram of the augmented soundscapes as inputs. This may demonstrate the effectiveness of the feature augmentation technique in emulating soundscape augmentation in the feature space instead of the waveform space.

[0159] FIG. 15A is a plot of ISO Pleasantness as a function of log gain illustrating gain interpolation using an additive attention (AA) with CONV augmentation (seed 5, validation fold 2) for masker class “water” according to various embodiments. FIG. 15B is a plot of ISO Pleasantness as a function of log gain illustrating gain interpolation using an additive attention (AA) with CONV augmentation (seed 5, validation fold 2) for masker class “traffic” according to various embodiments. FIG. 15C is a plot of ISO Pleasantness as a function of log gain illustrating gain interpolation using an additive attention (AA) with CONV augmentation (seed 5, validation fold 2) for masker class “bird” according to various embodiments. FIG. 15D is a plot of ISO Pleasantness as a function of log gain illustrating gain interpolation using an additive attention (AA) with CONV augmentation (seed 5, validation fold 2) for masker class “construction” according to various embodiments. FIG. 15E is a plot of ISO Pleasantness as a function of log gain illustrating gain interpolation using an additive attention (AA) with CONV augmentation (seed 5, validation fold 2) for masker class “silence” according to various

embodiments. All maskers are adjusted to 65.0 dBA at zero log-gain for the purpose of this visualization. The base soundscape has an in-situ SPL level of 65.32 dBA. The base soundscape and all maskers are unseen by the model during training. The solid lines in FIGS. 15A-E represent the means of the predicted distributions. The shaded regions represent the pleasantness levels up to one predicted standard deviation above and below the predicted means.

[0160] FIGS. 15A-E may demonstrate the gain-aware nature of the model by querying the ISO pleasantness distribution across 256 log-gain values in $[-2, 2]$ for different types of maskers. A silent ‘masker’ track is also included to test the model’s ability to consider in tandem the effects of both the masker and its gain level. Although not perfectly constant, it can be seen that the model can discern when the masker is silent and output a roughly constant pleasantness prediction regardless of the querying gain value. For other non-silent maskers, it can be seen that the model is aware of the nature of each masker. In line with the findings in previous works, the model predicted increasing pleasantness as the gain of the bird masker increases until it is close to the ambience level ($\gamma \approx -0.5$), followed by a drop in pleasantness below that of the unaugmented soundscape (i.e., the silent masker prediction). A similar effect is also seen with the water masker, with the pleasantness level being above that of the unaugmented soundscape up to $\gamma \approx -0.5$, although the pleasantness level starts dropping at a lower gain than the bird masker due to the more continuous nature of the water sound. For maskers that are known to be unpleasant, such as traffic and construction, the model correctly outputs increasingly unpleasant predictions as the gain level increases.

[0161] Conclusions

[0162] Various embodiments may relate to an improved probabilistic perceptual attribute predictor model that allows gain-aware prediction of subjective responses to augmented soundscapes. The proposed model decouples the masker and soundscape feature extraction and

emulates the soundscape augmentation process in the feature space, thereby eliminating the need for the computationally expensive mixing process in the waveform domain. Additionally, the model was reformulated to consider the digital gain level of the masker instead of the soundscape-to-masker ratio, allowing the use of maskers from various sources without the need for time-consuming calibration. The modular design of the model allows for significant feature reuse and pre-computation during inference time, reducing the overall latency and computational resources required in deployment. Using a large-scale dataset of 18K subjective responses from 442 participants, the ability of the model to accurately predict the pleasantness score in relation to the soundscape, masker and gain has been demonstrated.

[0163] The selection of maskers and playback gain levels in a soundscape augmentation system may be crucial to its effectiveness in improving the overall acoustic comfort of a given environment. Traditionally, the selection of appropriate maskers and gain levels has been informed by expert opinion, which may not be representative of the target population, or by listening tests, which can be time-consuming and labour-intensive. Furthermore, the resulting static choices of masker and gain are often inflexible to the dynamic nature of real-world soundscapes. In this work, a deep learning model is used to perform joint selection of the optimal masker and its gain level for a given soundscape. The proposed model was designed with highly modular building blocks, allowing for an optimized inference process that can quickly search through a large number of masker and gain combinations. In addition, the use of feature-domain soundscape augmentation conditioned on the digital gain level is introduced, eliminating the computationally expensive waveform domain mixing process during inference time, as well as the tedious pre-calibration process required for maskers. The proposed system was validated on a large-scale dataset of subjective responses to augmented soundscapes with more than 440 participants, ensuring the ability of the model to predict the combined effect of the masker and its gain level on the perceptual pleasantness level.

Claims

1. A soundscape augmentation system comprising:
 - a data acquisition system configured to provide ambient soundscape data;
 - a database including a plurality of masker configurations;
 - a perceptual attribute predictor coupled to the data acquisition system and the database, the perceptual attribute predictor configured to generate predictions representing perception on one or more pre-defined perceptual attribute scales for each masker configuration of the plurality of masker configurations based on the ambient soundscape data;
 - a masker configuration ranking system configured to determine one or more optimal masker configurations based on the predictions generated by the perceptual attribute predictor; and
 - a playback system configured to play back or reproduce the one or more optimal masker configurations.
2. The soundscape augmentation system according to claim 1,
 - wherein the data acquisition system is also configured to provide demographic or contextual data.
3. The soundscape augmentation system according to claim 2,
 - wherein the perceptual attribute predictor is configured to generate the predictions also based on the demographic or contextual data.
4. The soundscape augmentation system according to any one of claims 2 to 3,
 - wherein the demographic or contextual data comprises data related to environmental parameters, demographic data of one or more listeners, psychological data of the one or more listeners, physiological data of the one or more listeners, or any combination thereof.
5. The soundscape augmentation system according to claim 4,

wherein the data related to environmental parameters comprises data related to location, data related to the visual environment, number of people nearby, air temperature, humidity, wind speed, or any combination thereof.

6. The soundscape augmentation system according to claim 4 or claim 5,
wherein the demographic data of the one or more listeners comprises data related to age, gender, occupation, or any combination thereof.
7. The soundscape augmentation system according to any one of claims 4 to 6,
wherein the psychological data of the one or more listeners comprises data related to noise sensitivity, perceived stress, well-being index scores, or any combination thereof.
8. The soundscape augmentation system according to any one of claims 4 to 7,
wherein the physiological data of the one or more listeners comprises data related to heart rate, blood pressure, body temperature, or any combination thereof.
9. The soundscape augmentation system according to any one of claims 1 to 8,
wherein the perceptual attribute predictor is configured to generate predictions based on one or more masker gain inputs.
10. The soundscape augmentation system according to any one of claims 1 to 9,
wherein the one or more pre-defined perceptual attribute scales comprise pleasantness, vibrancy, eventfulness, calmness, perceived loudness, sound quality, sharpness, roughness, or any combination thereof.
11. The soundscape augmentation system according to any one of claims 1 to 10,
wherein the perceptual attribute predictor is configured to generate the predictions by combining the ambient soundscape data and each masker configuration.

12. The soundscape augmentation system according to any one of claims 1 to 11,
wherein the perceptual attribute predictor comprises a soundscape feature extractor configured to extract features of the ambient soundscape data;
wherein the perceptual attribute predictor comprises a masker feature extractor configured to extract features of each masker configuration; and
wherein the perceptual attribute predictor is configured to generate the predictions based on the extracted features of the ambient soundscape data and the extracted features of each masker configuration.
13. The soundscape augmentation system according to any one of claims 1 to 12,
wherein the perceptual attribute predictor comprises an objective feature block configured to extract objective features from the ambient soundscape data and each masker configuration;
wherein the perceptual attribute predictor comprises a subjective feature block configured to extract subjective features from the demographic or contextual data; and
wherein the perceptual attribute predictor is configured to generate the predictions based on the extracted objective features and the extracted subjective features.
14. The soundscape augmentation system according to any one of claims 1 to 13,
wherein the perceptual attribute predictor comprises linear regression models configured to generate the predictions based on acoustic or psychoacoustic parameters computed based on the ambient soundscape data and each masker configuration.
15. The soundscape augmentation system according to any one of claims 1 to 14,
wherein the perceptual attribute predictor comprises one or more deep neural networks configured to generate the predictions based on raw audio, spectrogram representations computed based on the ambient soundscape data and each masker configuration, or a combination of the raw audio and the spectrogram representations.

16. The soundscape augmentation system according to any one of claims 1 to 15,
wherein the data acquisition system is configured to generate the ambient soundscape data based on inputs from one or more microphones, one or more antennas, a storage medium, or any combination thereof.
17. The soundscape augmentation system according to any one of claims 1 to 16,
wherein the plurality of masker configurations comprises one or more audio tracks of recorded or synthesized sounds, one or more audio tracks of silence, or one or more audio tracks derived from the one or more audio tracks of recorded or synthesized sounds and the one or more audio tracks of silence.
18. The soundscape augmentation system according to any one of claims 1 to 17,
wherein the playback system comprises one or more speakers, one or more virtual or augmented reality headsets, one or more headphones, one or more earbuds, or any combination thereof.
19. The soundscape augmentation system according to any one of claims 1 to 18,
wherein the predictions generated are non-deterministic.
20. A method of forming a soundscape augmentation system, the method comprising:
providing a data acquisition system configured to provide ambient soundscape data;
providing a database including a plurality of masker configurations;
coupling a perceptual attribute predictor to the data acquisition system and the database, the perceptual attribute predictor configured to generate predictions representing perception on one or more pre-defined perceptual attribute scales for each masker configuration of the plurality of masker configurations based on the ambient soundscape data;
providing a masker configuration ranking system configured to determine one or more optimal masker configurations based on the predictions generated by the perceptual attribute predictor; and
providing a playback system configured to play back or reproduce the one or more optimal masker configurations.

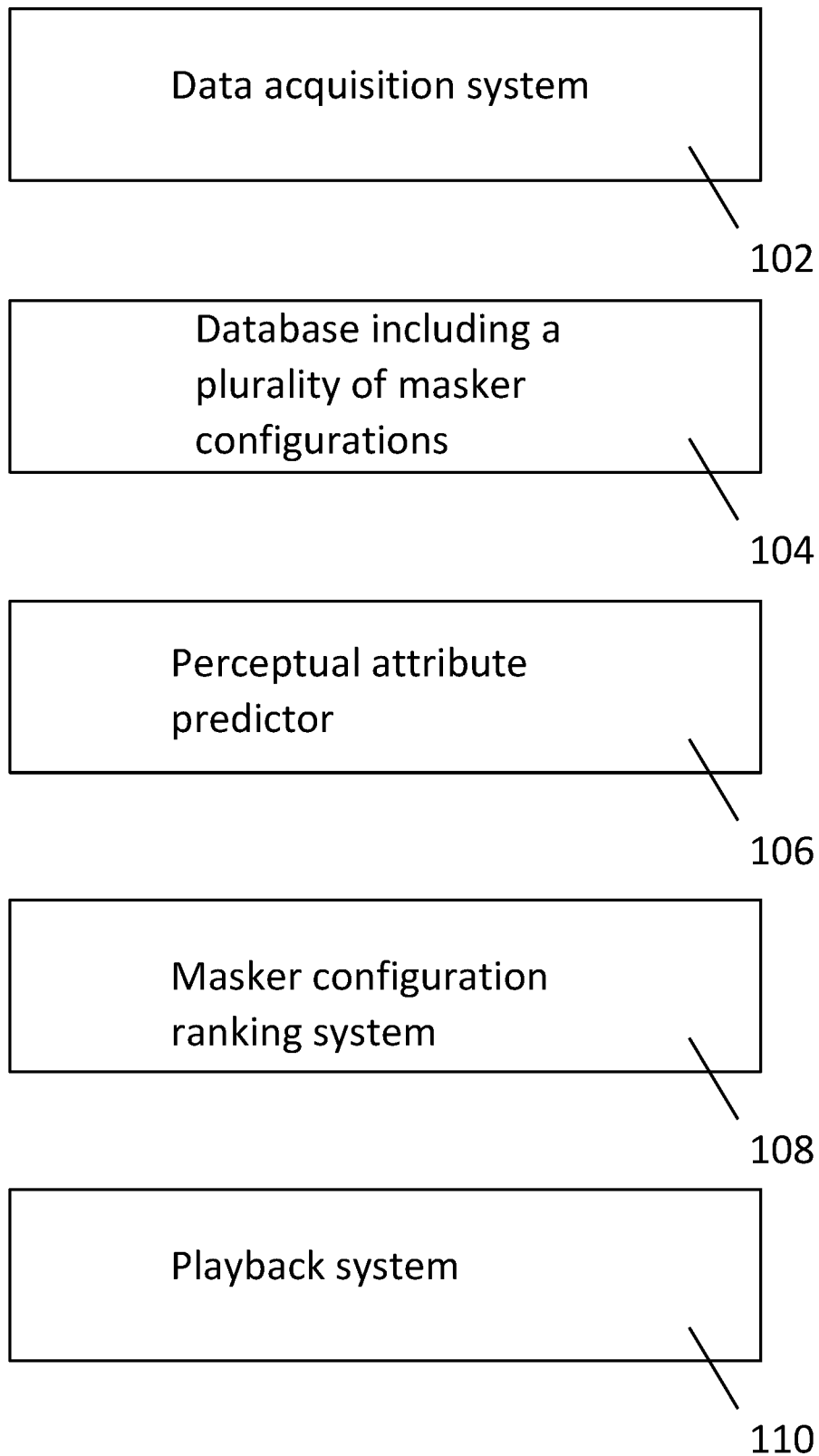
FIG. 1

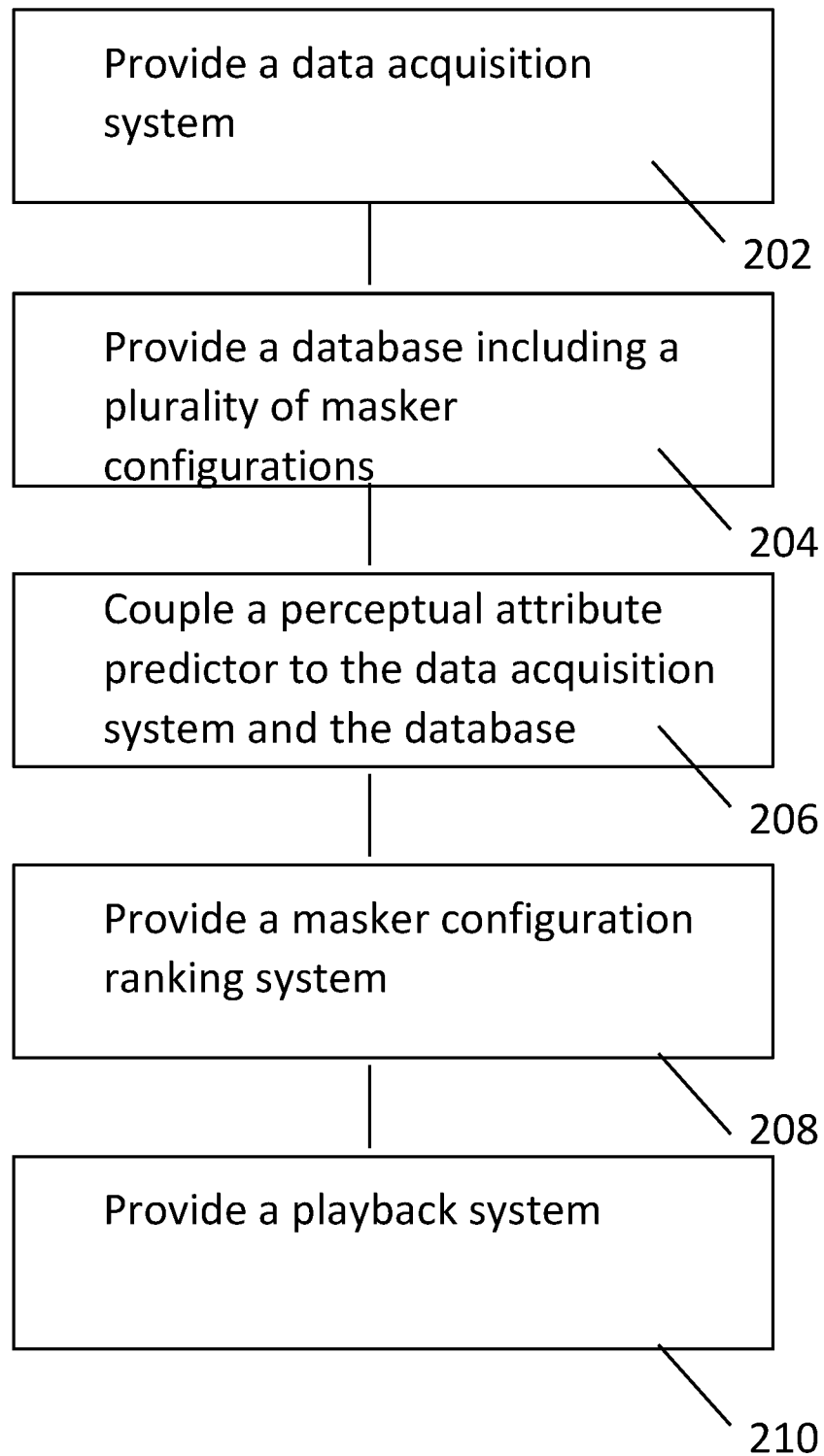
FIG. 2

FIG. 3

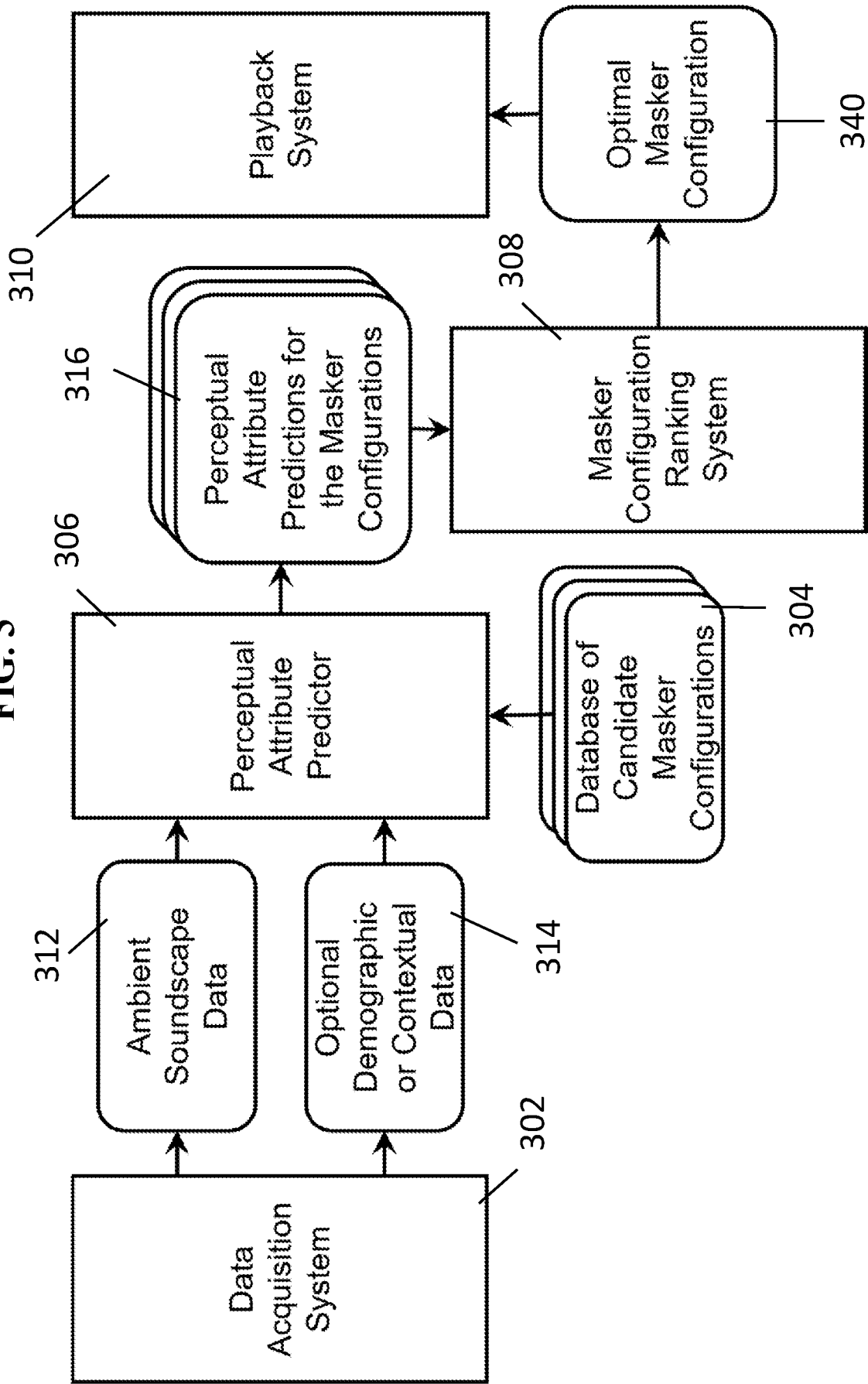


FIG. 4

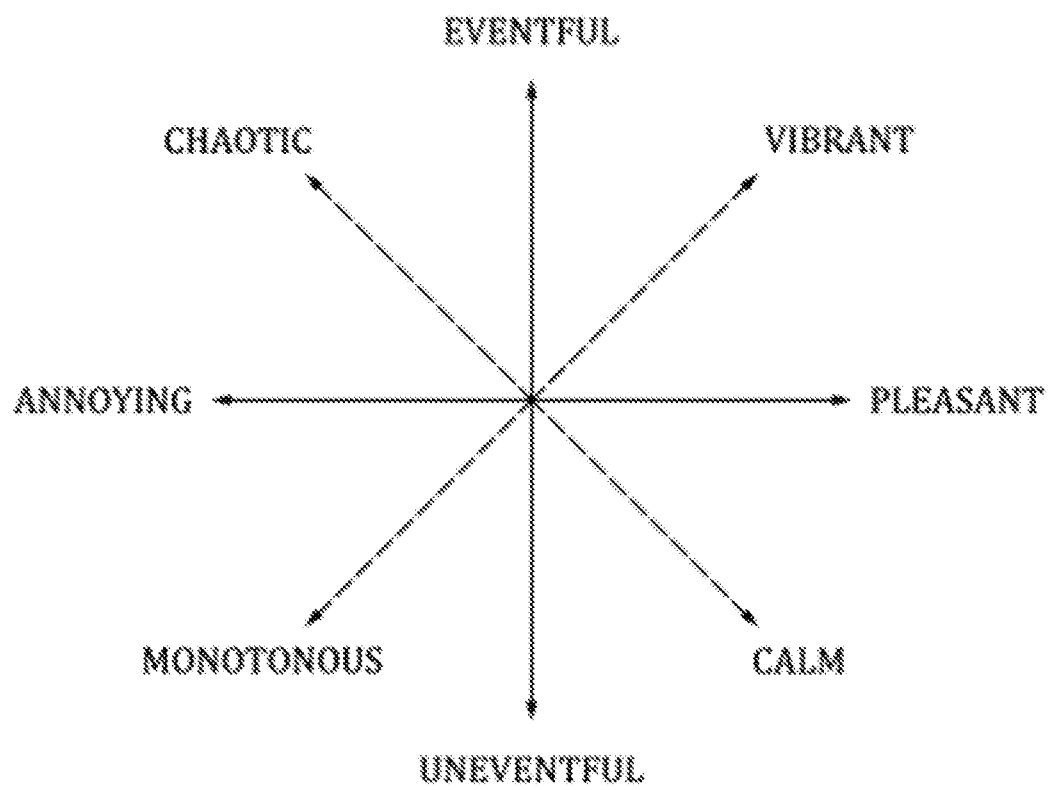


FIG. 5

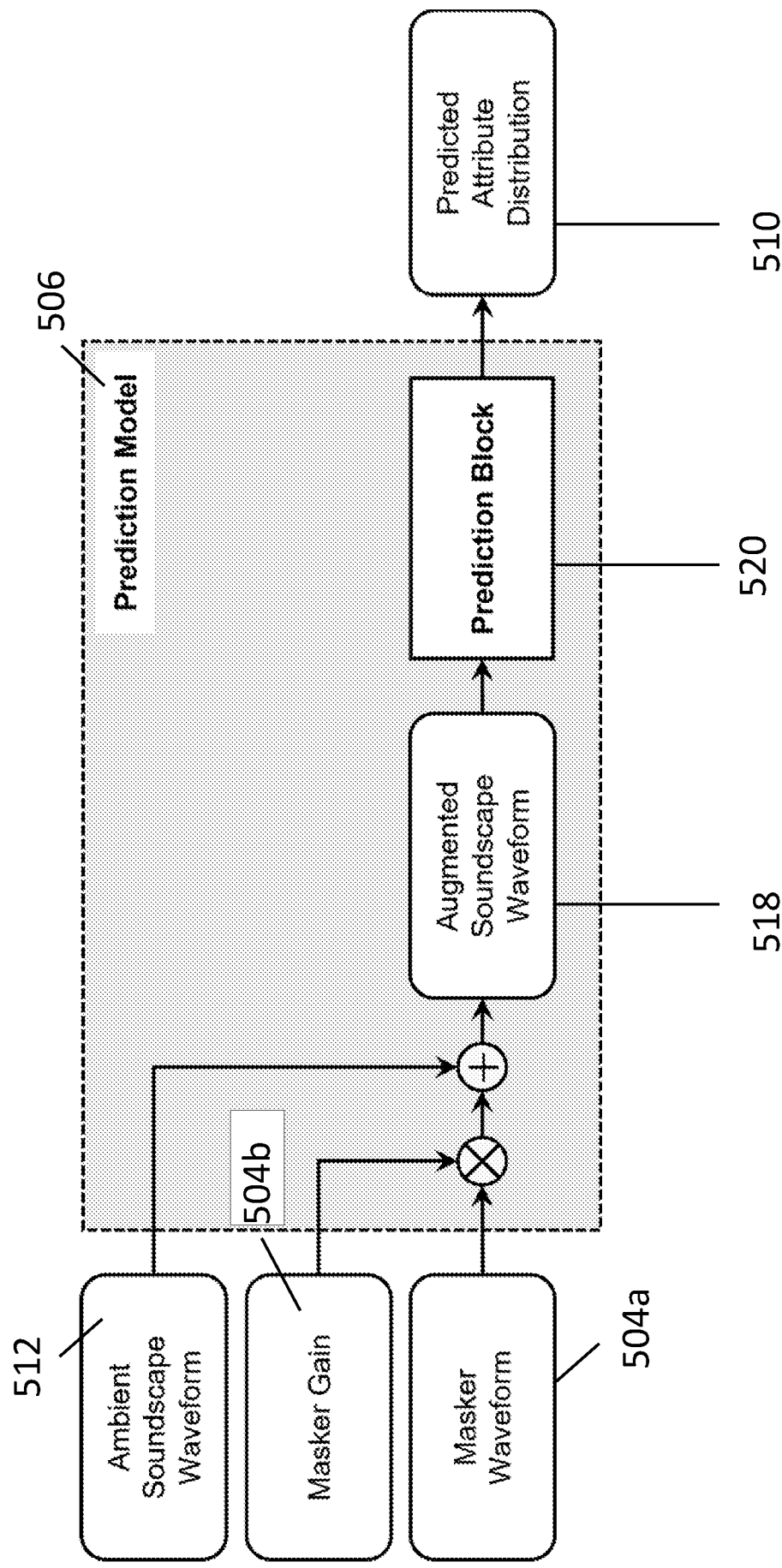


FIG. 6

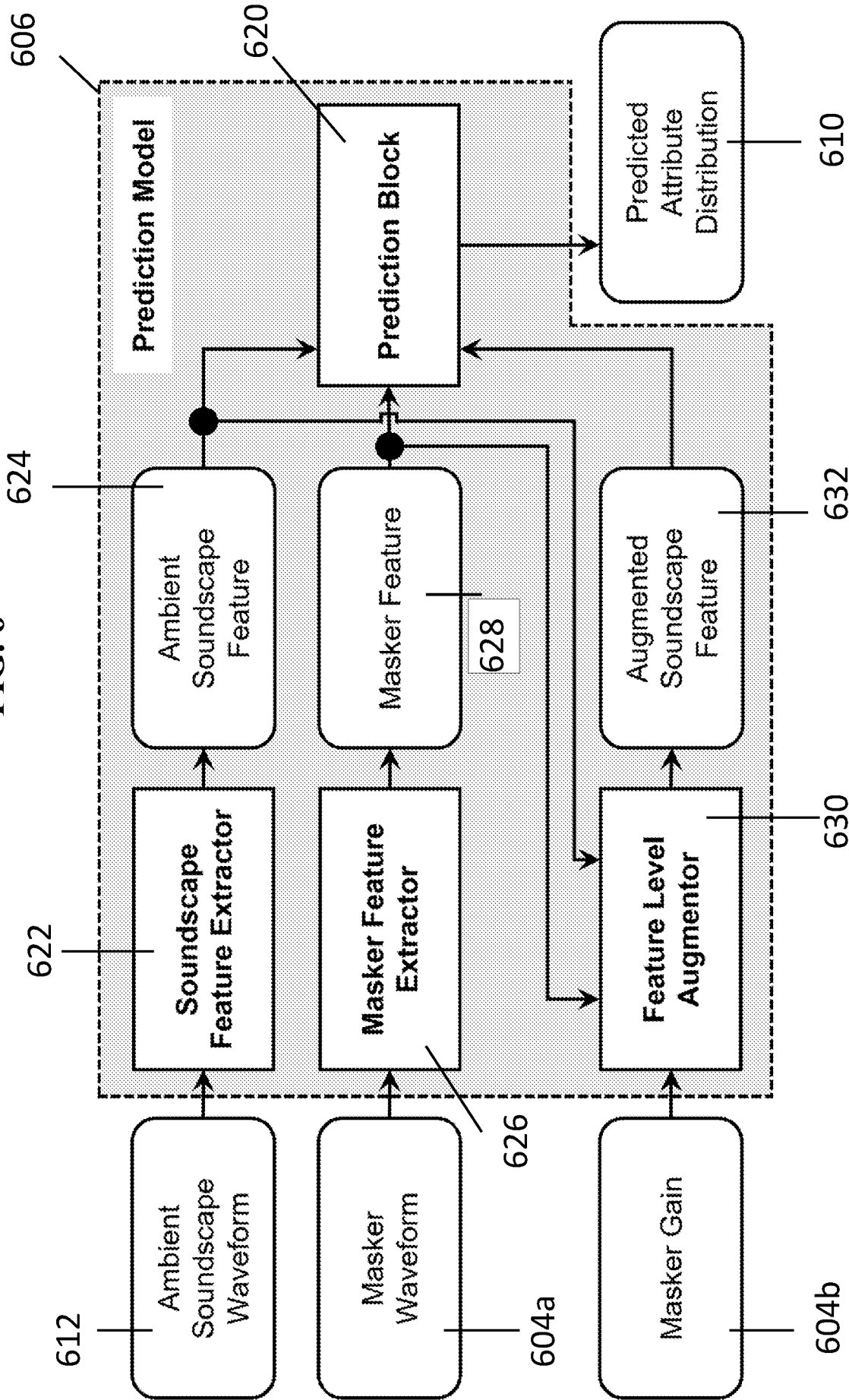


FIG. 7

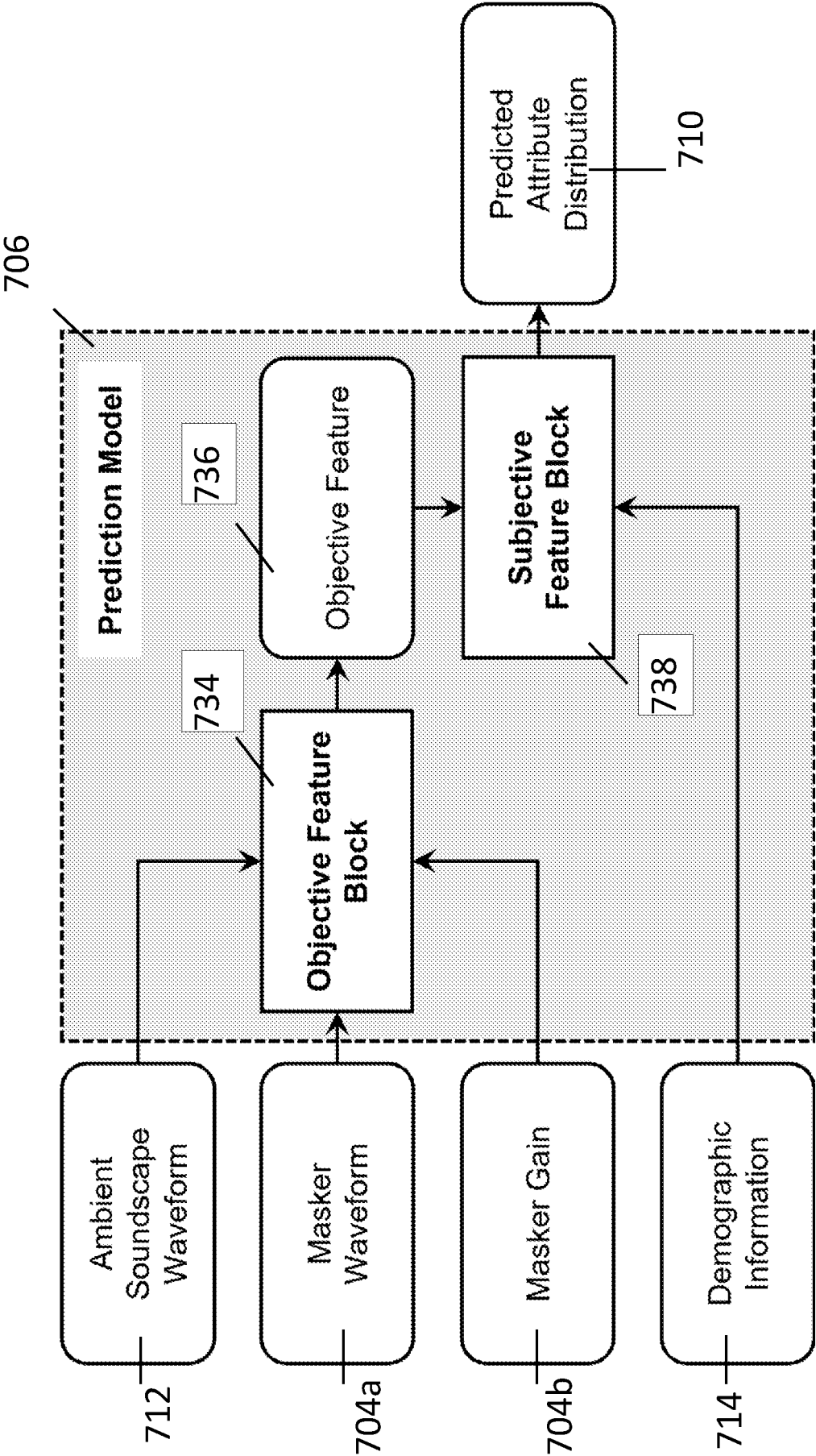


FIG. 8

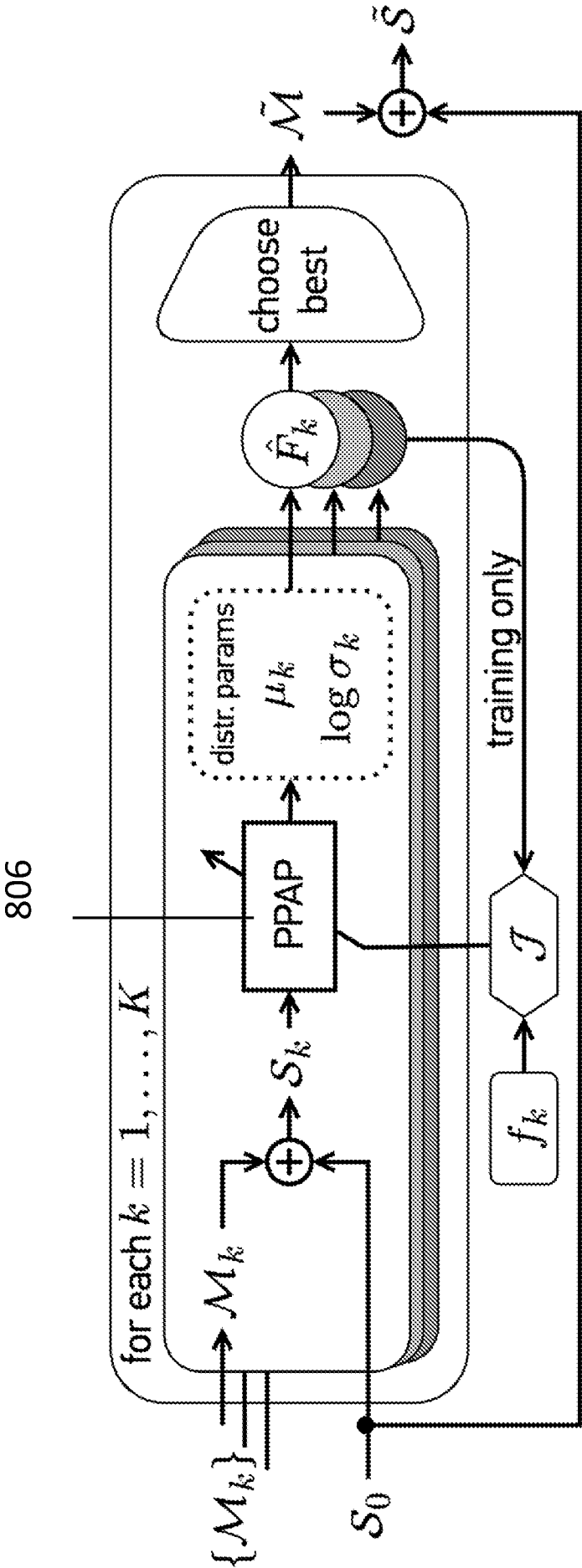


FIG. 9

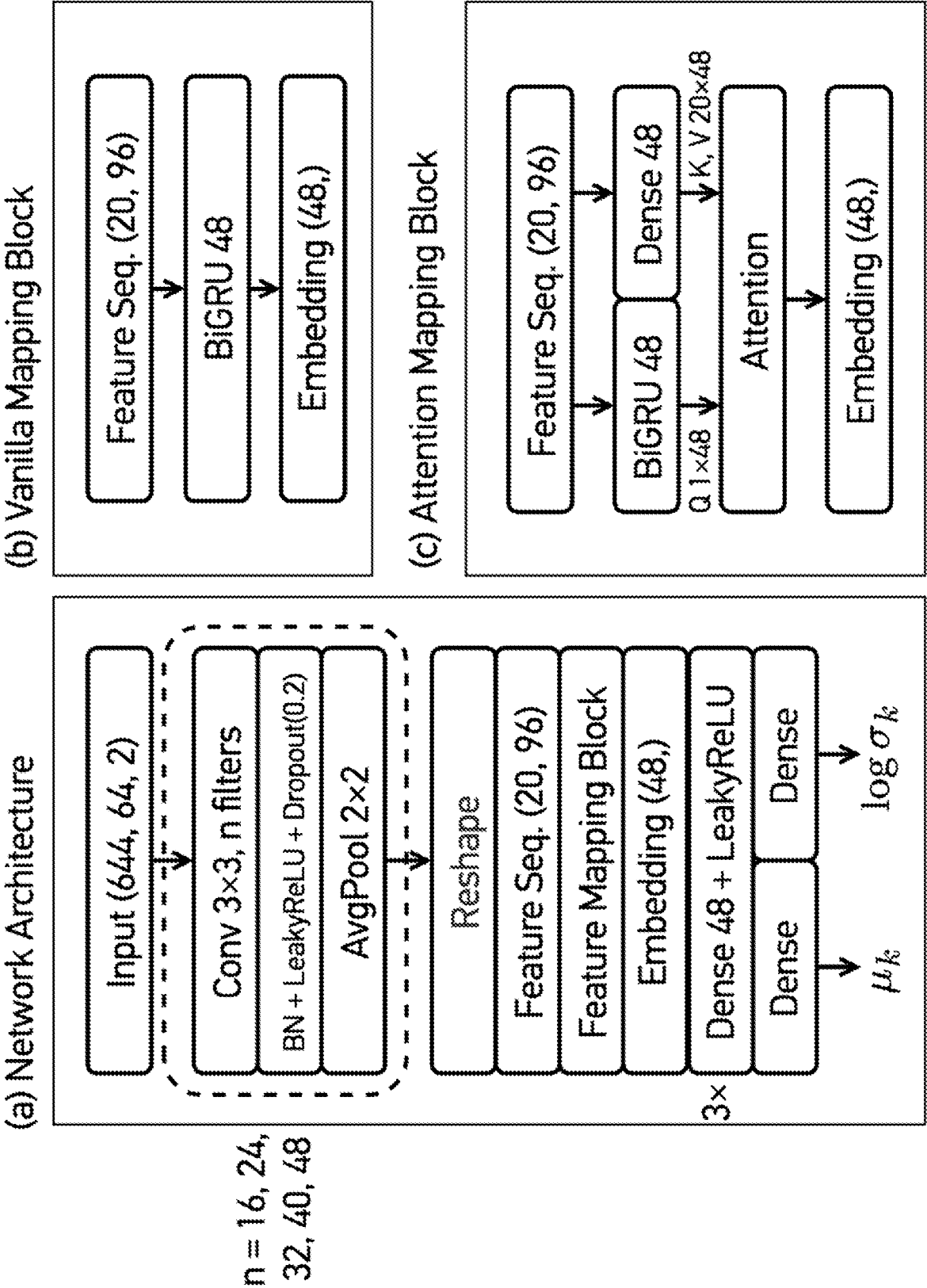


FIG. 10

Cross-Validation Set				
Model	Params.	Deterministic	Probabilistic	Improv. (%)
(Label mean)	—	0.1556	—	—
PPAP w/o attention	76K	0.1262±0.0009	0.1255±0.0011	0.5
PPAP w/ additive att.	85K	0.1264±0.0015	0.1251±0.0014	1.0*
PPAP w/ dot-prod. att.	85K	0.1259±0.0012	0.1250±0.0013	0.7*
PPAP w/ multi-head att.	123K	0.1244±0.0009	0.1230±0.0010	1.1*

Test Set				
Model	Params.	Deterministic	Probabilistic	Improv. (%)
(Label mean)	—	0.1181	—	—
PPAP w/o attention	76K	0.1106±0.0114	0.1020±0.0096	7.8*
PPAP w/ additive att.	85K	0.1030±0.0137	0.0970±0.0080	5.8*
PPAP w/ dot-prod. att.	85K	0.0950±0.0131	0.0906±0.0091	4.6
PPAP w/ multi-head att.	123K	0.0999±0.0147	0.0996±0.0104	0.3

FIG. 11

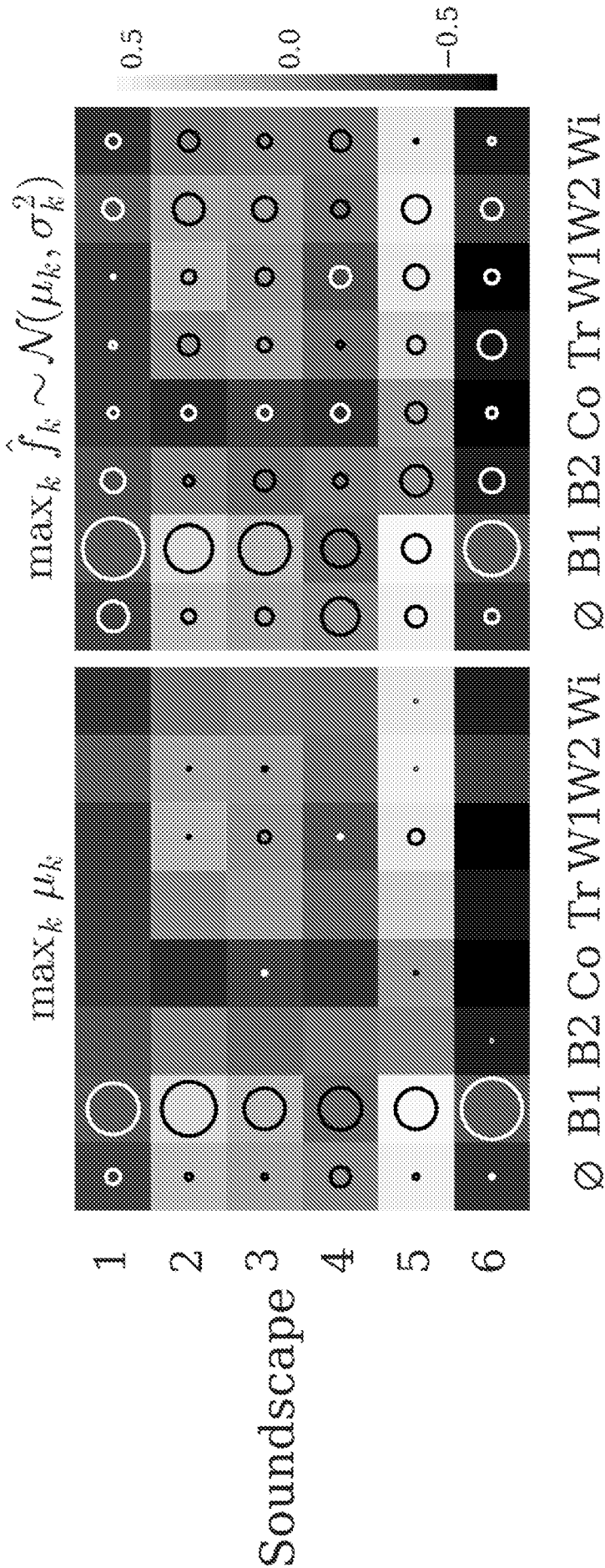


FIG. 12A

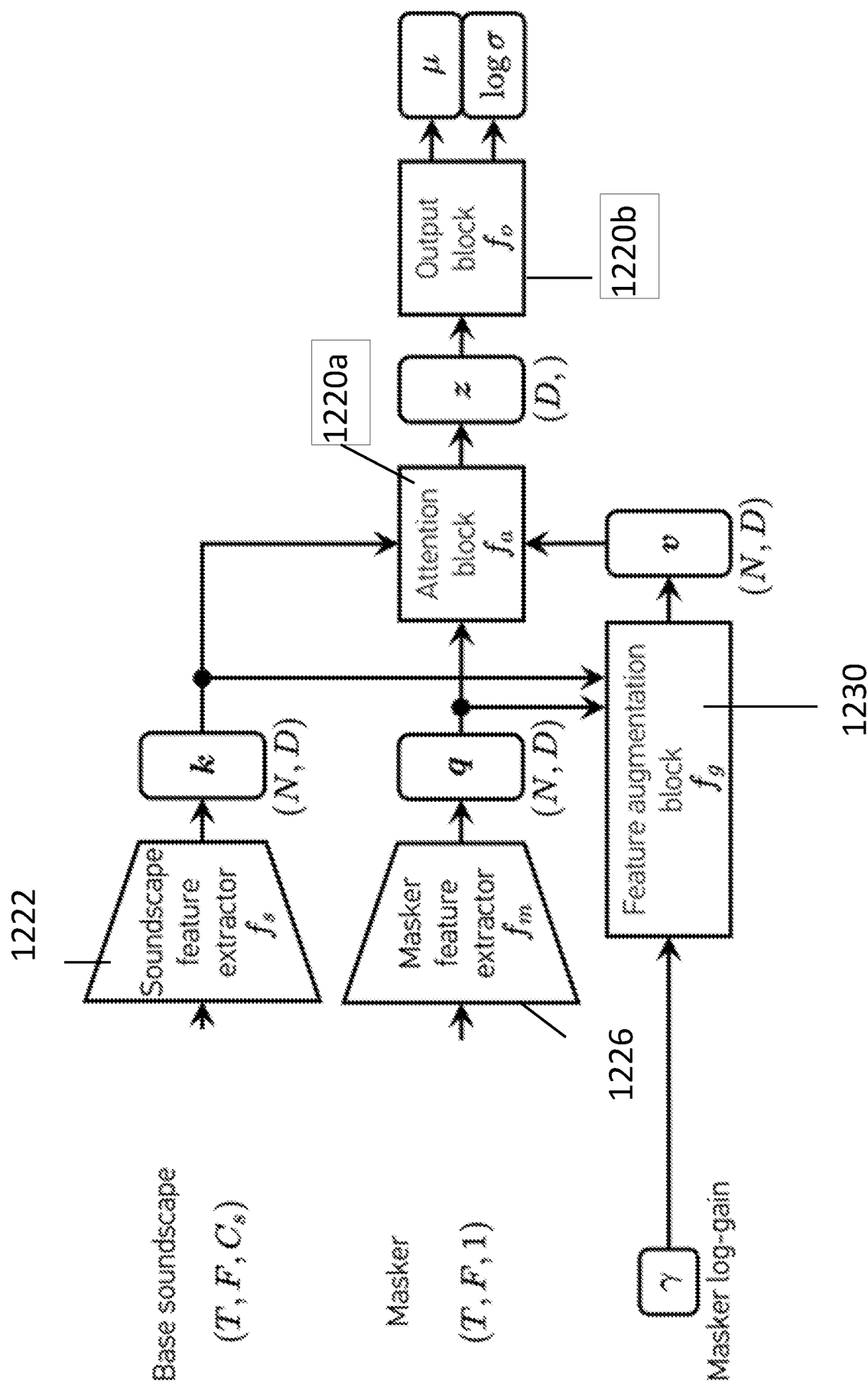


FIG. 12B

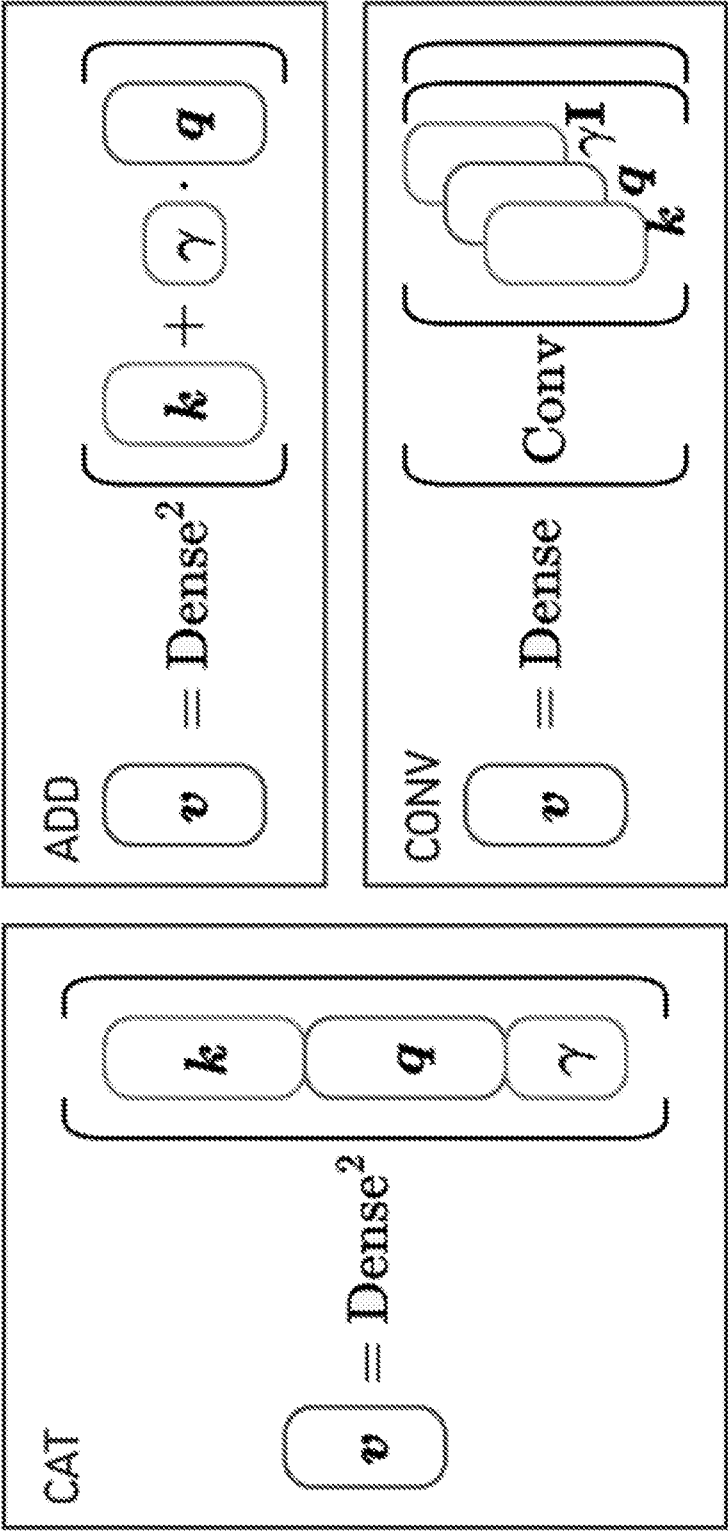


FIG. 13

Algorithm 1: Optimized Inference**Input:** Soundscape recording s_i 1 $S_i = \text{LogMelSpec}(s_i)$ 2 $k_i = f_s(S_i)$ 3 **foreach** masker spectrogram M_j **do**4 $q_j = f_m(M_j)$ // can be precomputed5 **foreach** gain g **do**6 $v_{i,j,g} = f_g(k_i, q_j, \gamma)$ 7 $Y_{i,j,g} = f_o \circ f_a(q_j, k_i, v_{i,j,g})$ 8 **end**9 **end****Output:** Predicted pleasantness distributions $\{Y_{i,j,g}\}$

FIG. 14

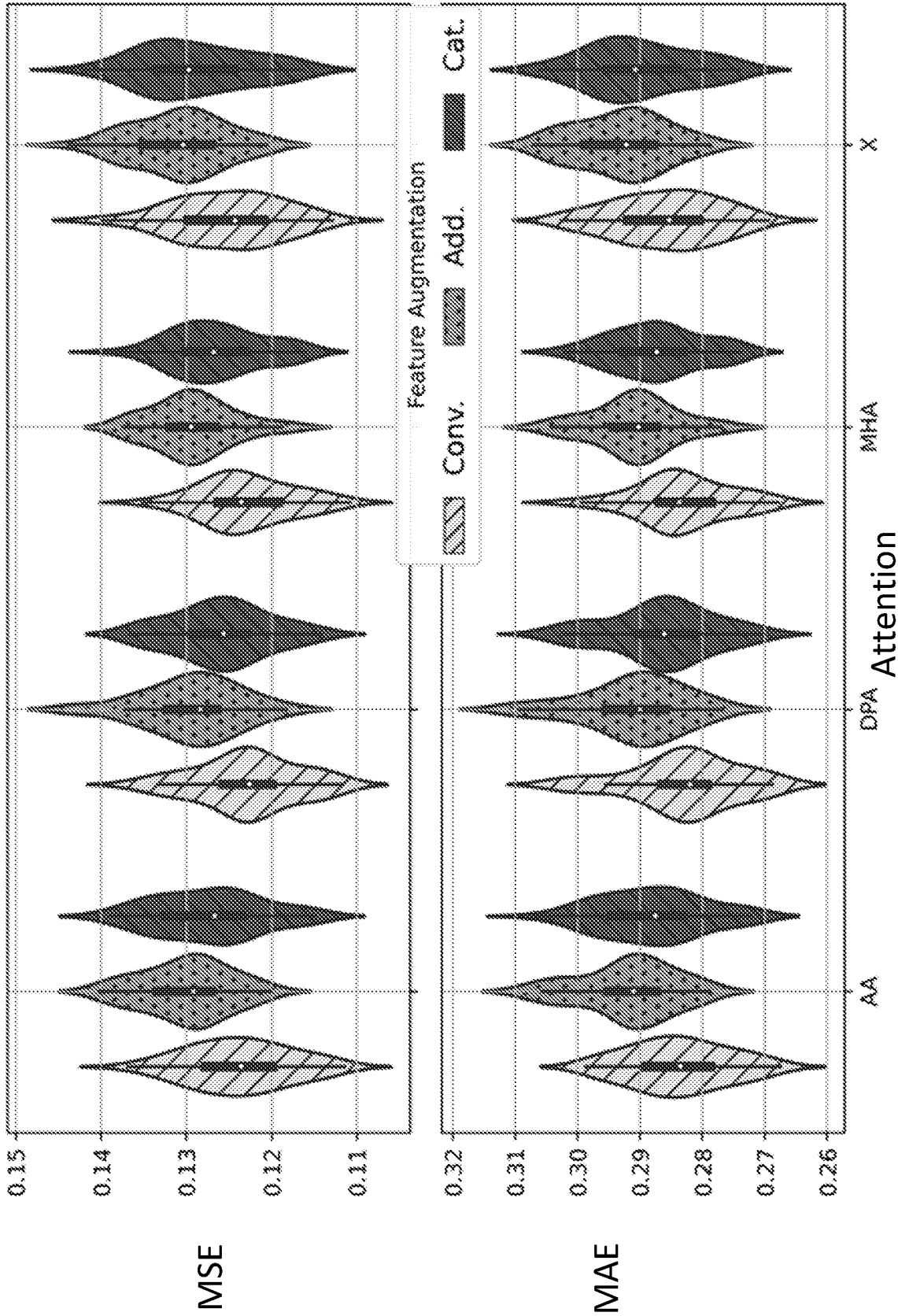


FIG. 15A

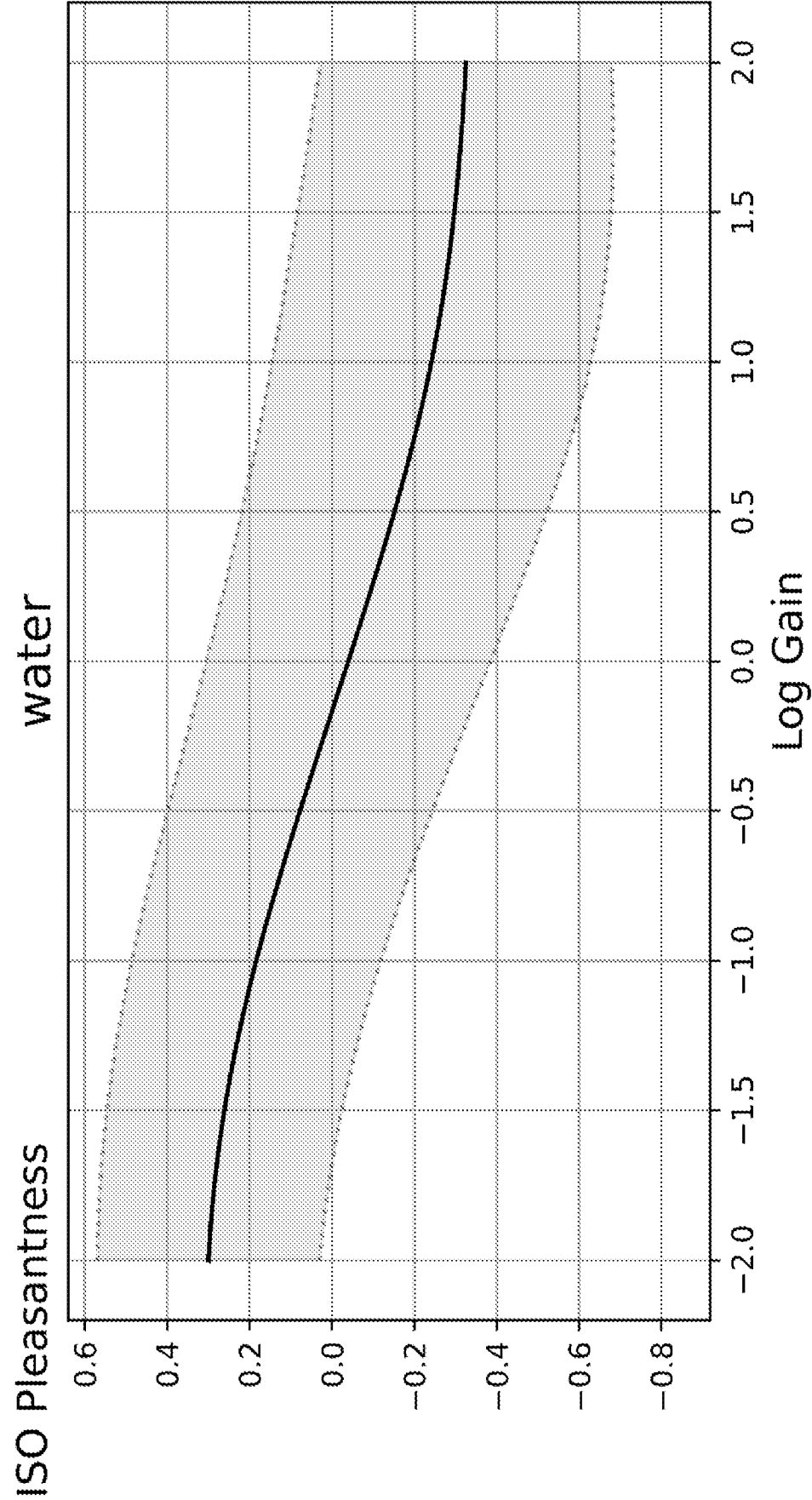


FIG. 15B

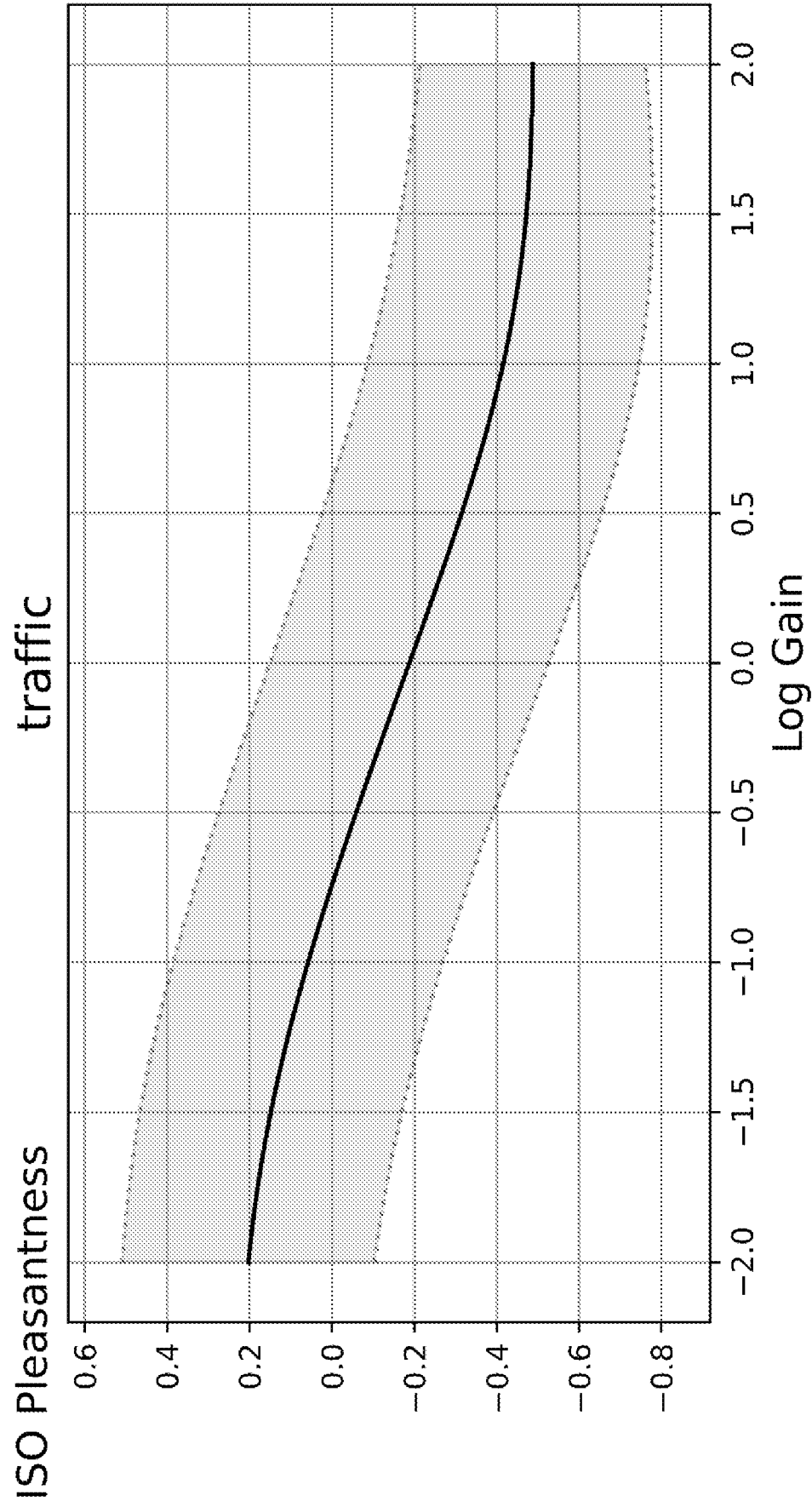


FIG. 15C

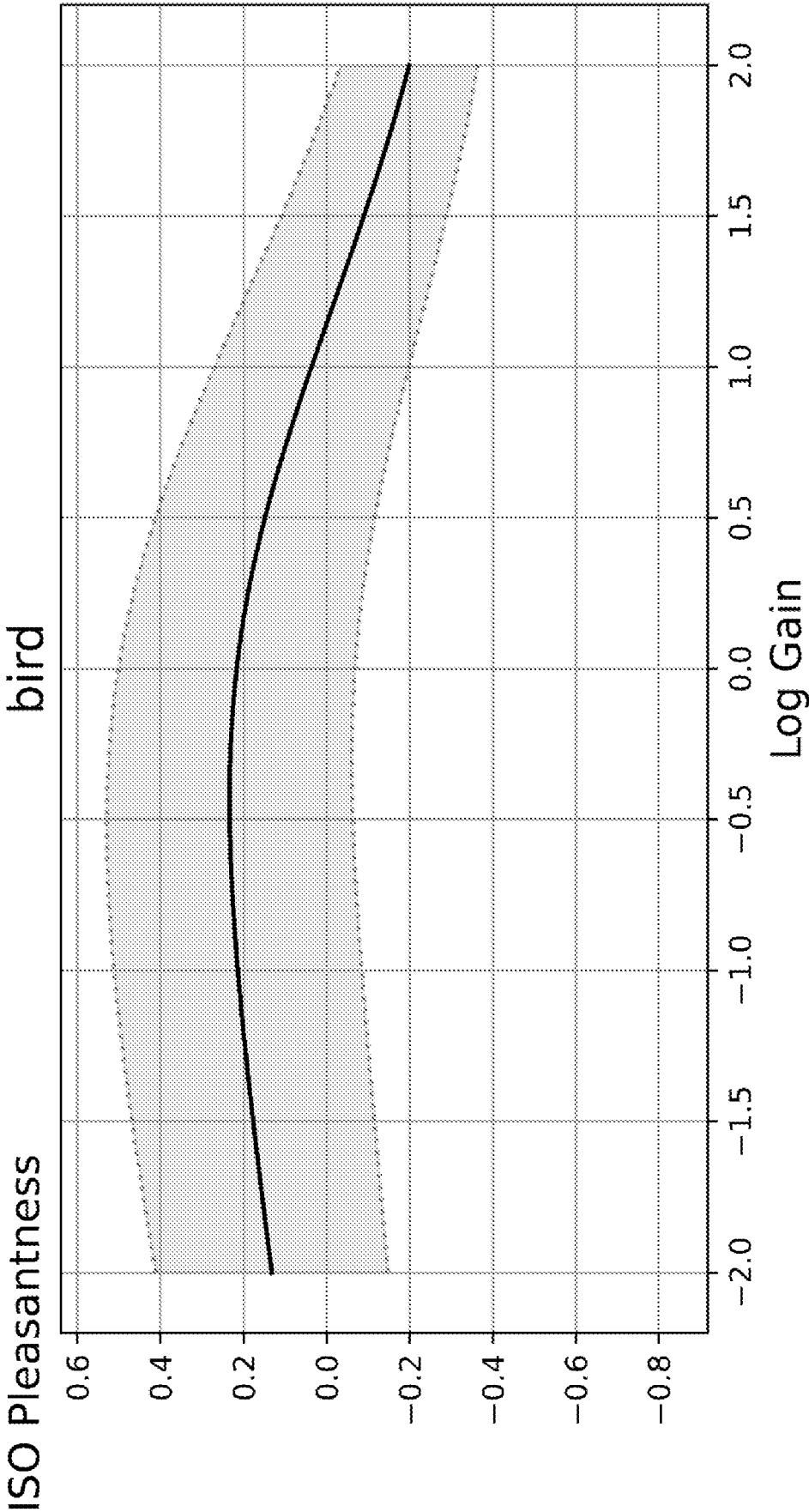


FIG. 15D

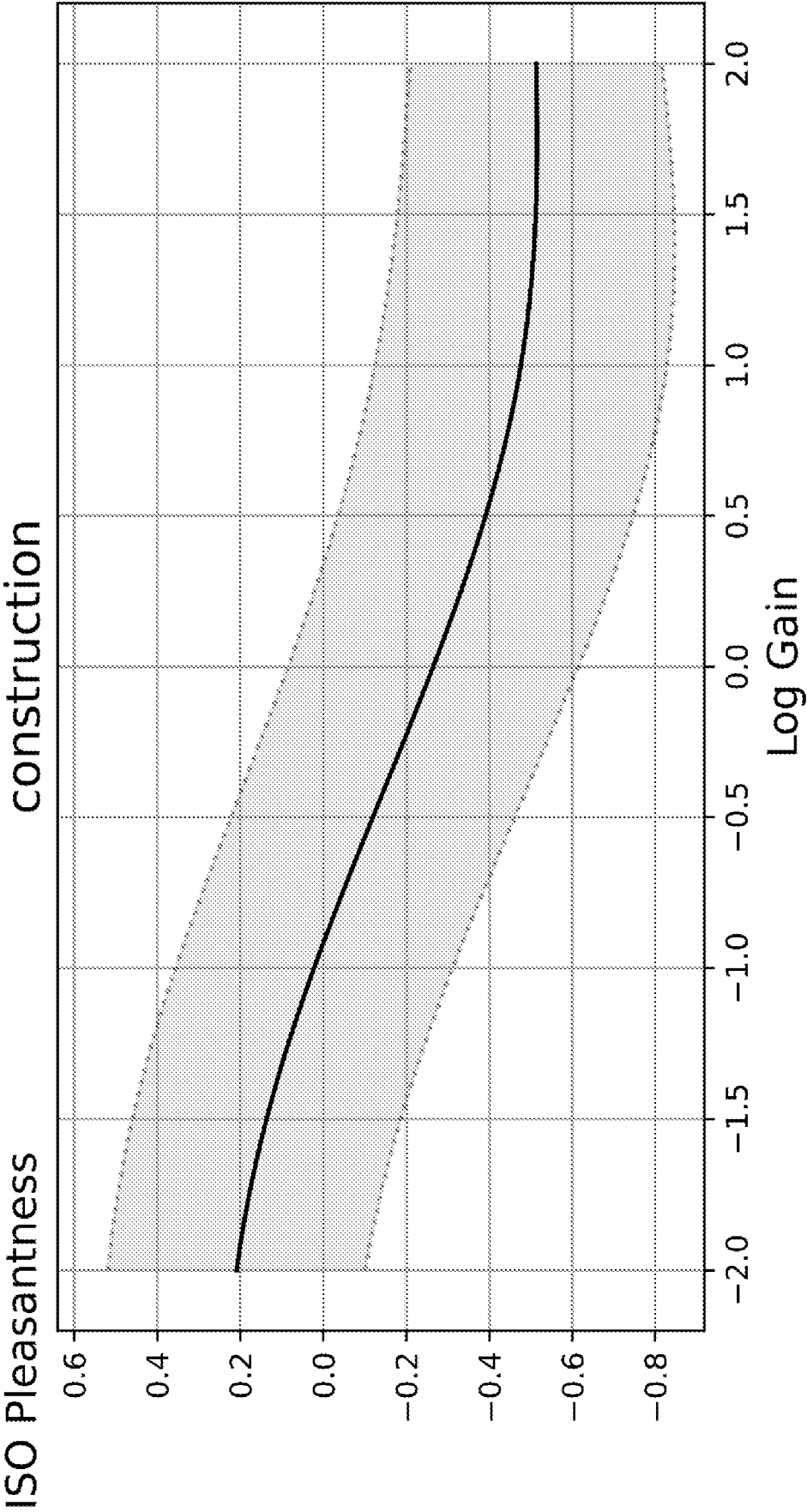
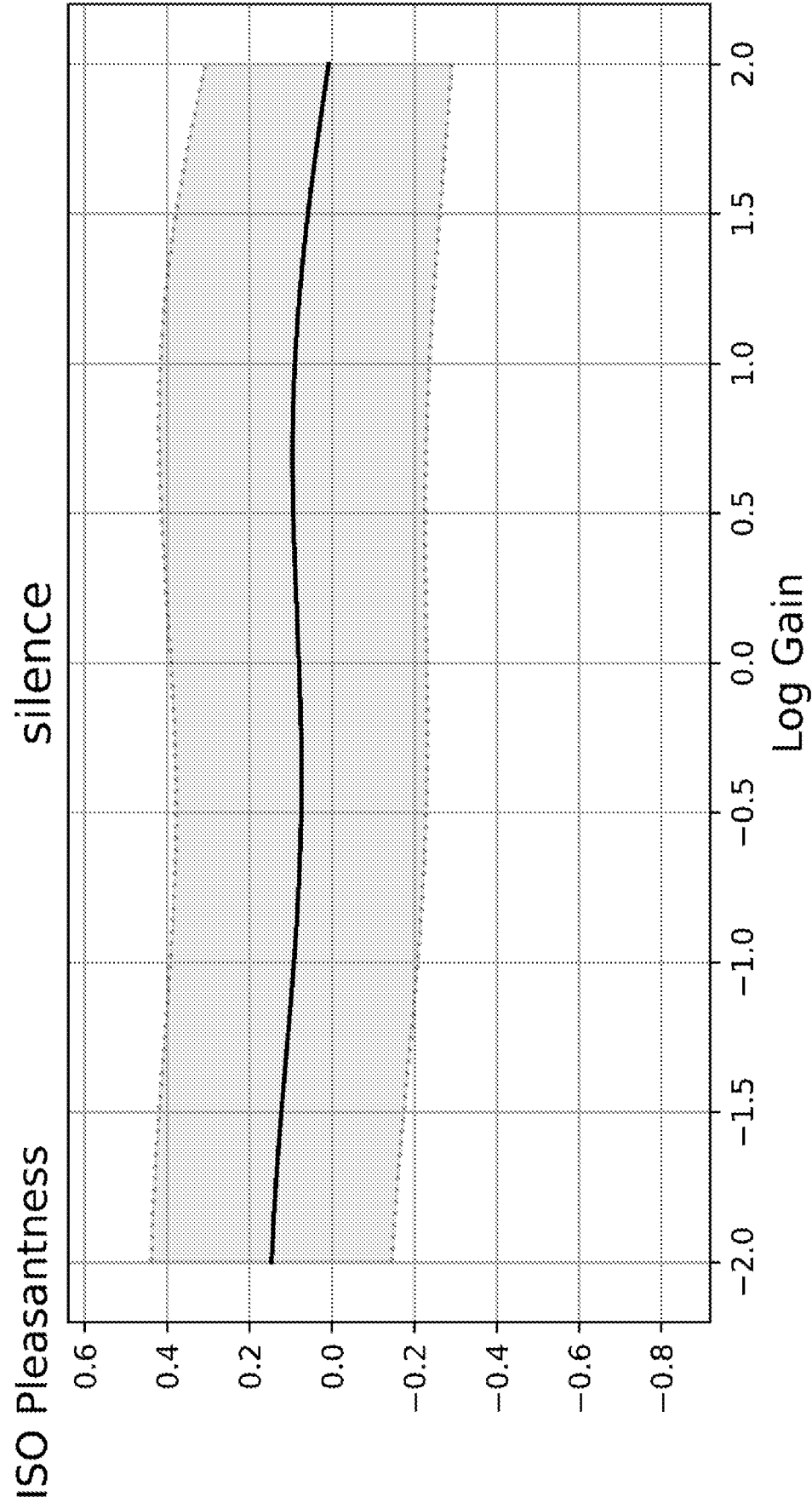


FIG. 15E



INTERNATIONAL SEARCH REPORT

International application No.

PCT/SG2023/050289

A. CLASSIFICATION OF SUBJECT MATTER

G10K 11/175 (2006.01)

According to International Patent Classification (IPC)

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10K

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)

FAMPAT, Internet: augment, improve, attenuate, soundscape, sound mask, audio, acoustic, predict, estimate, forecast, perception, impression, reaction, perceive, mask and related terms.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2022/073128 A1 (MINDSET INNOVATION, INC.) 14 April 2022 Paragraphs [0056]-[0059], [0063]-[0071], [0077], [0079]-[0081], [0084]-[0085], [0102], [0111], [0115], [0137]-[0138], [0141]-[0142], [0152]-[0163], [0174]-[0180], [0194], [0205], [0265], [0308]-[0309]	1-20
X	US 2019/0013003 A1 (BAUGHMAN A. K. ET AL.) 10 January 2019 Paragraphs [0015], [0018]-[0019], [0021]-[0029], [0032], [0042]-[0043]	1-20
X	US 2021/0003980 A1 (STAVITSKII O. ET AL.) 7 January 2021 Paragraphs [0024], [0030]-[0031], [0035]-[0037], [0041]-[0048]	1-20
X	CN 110362789 A (SHANGHAI ACADEMY ENV SCIENCES) 22 October 2019 Paragraphs [0047]-[0052], [0054]-[0055] of the original non-English language document (a machine translation is enclosed only for your reference)	1-20
X	CN 109238448 A (SHANGHAI ACADEMY ENV SCIENCES) 18 January 2019 Paragraphs [0032]-[0036] of the original non-English language document (a machine translation is enclosed only for your reference)	1

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

*Special categories of cited documents:

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

"D" document cited by the applicant in the international application

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search

03/10/2023

(day/month/year)

Date of mailing of the international search report

06/10/2023

(day/month/year)

Name and mailing address of the ISA/SG



Intellectual Property Office of Singapore

1 Paya Lebar Link, #11-03

PLQ 1, Paya Lebar Quarter

Singapore 408533

Email: pct@ippos.gov.sg

Authorized officer

Lai Tan Lei Tanny (Dr)

IPOS Customer Service Tel. No.: (+65) 6339 8616

INTERNATIONAL SEARCH REPORT

International application No.

PCT/SG2023/050289

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	HONG J. Y. ET AL., Effects of adding natural sounds to urban noises on the perceived loudness of noise and soundscape quality. <i>Science of the Total Environment</i> , 20 November 2019, Vol. 711, pages 134571 [Retrieved on 2023-10-03] <DOI: 10.1016/J.SCITOTENV.2019.134571> Section 2.2 first, second and fifth paragraphs, section 2.3 second paragraph, section 2.4 first paragraph, section 4.3 first paragraph	1-20
A	OOI K. ET AL., A deep learning approach for modelling perceptual attributes of soundscapes. <i>INTER-NOISE and NOISE-CON Congress and Conference Proceedings</i> , 12 October 2020, Vol. 261, No. 3, pages 3580-3589 [Retrieved on 2023-10-03] <DOI: NONE> Section 2.1 fourth paragraph, section 2.3	2-8, 14, 15
A	US 2018/0364966 A1 (VALERI F. C. ET AL.) 20 December 2018 Paragraphs [0049], [0051], [0060]	14, 15
P,X	OOI K. ET AL., Probably Pleasant? A Neural-Probabilistic Approach to Automatic Masker Selection for Urban Soundscape Augmentation. <i>ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)</i> , 23 May 2022, pages 8887-8891 [Retrieved on 2023-10-03] <DOI: 10.1109/ICASSP43922.2022.9746897> Section 3, section 4.1, section 4.2	1-20
P,X	OOI K. ET AL., ARAUS: A Large-Scale Dataset and Baseline Models of Affective Responses to Augmented Urban Soundscapes. <i>arXiv:2207.01078</i> , 6 March 2023 [Retrieved on 2023-10-03] <DOI: 10.48550/ARXIV.2207.01078> Section 3, section 5.1	1-20
P,X	WATCHARASUPAT K. N. ET AL., Autonomous In-Situ Soundscape Augmentation via Joint Selection of Masker and Gain. <i>arXiv:2204.13883</i> , 23 July 2022 [Retrieved on 2023-10-03] <DOI: 10.48550/ARXIV.2204.13883> Sections II & III	1-20
P,X	US 2022/0383849 A1 (BENEDETTO W. M.) 1 December 2022 Paragraphs [0026], [0032], [0039]-[0040], [0042]-[0043], [0048]-[0051], [0055]-[0056], [0059], [0065]	1-20

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/SG2023/050289

Note: This Annex lists known patent family members relating to the patent documents cited in this International Search Report. This Authority is in no way liable for these particulars which are merely given for the purpose of information.

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2022/073128 A1	14/04/2022	NONE	
US 2019/0013003 A1	10/01/2019	NONE	
US 2021/0003980 A1	07/01/2021	US 2020/0142371 A1 RU 2769275 C1 WO 2020/095105 A1 JP 2022506562 A EP 3877939 A1 CN 112970041 A KR 20210068570 A US 2022/0229408 A1	07/05/2020 29/03/2022 14/05/2020 17/01/2022 15/09/2021 15/06/2021 09/06/2021 21/07/2022
CN 110362789 A	22/10/2019	NONE	
CN 109238448 A	18/01/2019	NONE	
US 2018/0364966 A1	20/12/2018	CN 109086287 A DE 102018114039 A1	25/12/2018 20/12/2018
US 2022/0383849 A1	01/12/2022	WO 2022/251077 A1	01/12/2022



(21) 申请号 202380036570.3

(22) 申请日 2023.04.26

(30) 优先权数据

10202204451S 2022.04.27 SG

(85) PCT国际申请进入国家阶段日

2024.10.25

(86) PCT国际申请的申请数据

PCT/SG2023/050289 2023.04.26

(87) PCT国际申请的公布数据

W02023/211385 EN 2023.11.02

(71) 申请人 南洋理工大学

地址 新加坡

申请人 建屋发展局

(72) 发明人 黄文睿 黄株绮 林滨 王镇廷

黄智明 颜允圣

(74) 专利代理机构 中原信达知识产权代理有限公司 11219

专利代理师 韩峰 孙志湧

(51) Int.Cl.

G10K 11/175 (2006.01)

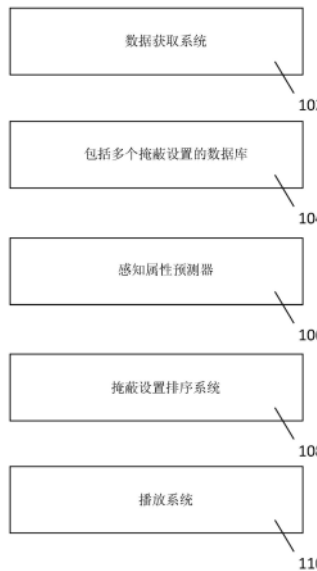
权利要求书2页 说明书24页 附图20页

(54) 发明名称

声景增强系统及其形成方法

(57) 摘要

各个实施例可以提供声景增强系统。声景增强系统可以包括数据获取系统,数据获取系统被设置为提供周围声景数据。声景增强系统还可以包括数据库,数据库包括多个掩蔽设置。声景增强系统还可以包括感知属性预测器,感知属性预测器耦接于数据获取系统和数据库,感知属性预测器被设置为基于周围声景数据,为多个掩蔽设置中的每个掩蔽设置生成在一个或多个预定义感知属性指标上代表感知的预测。声景增强系统可以附加地包括掩蔽设置排序系统,掩蔽设置排序系统被设置为基于感知属性预测器生成的预测,确定一个或多个最佳掩蔽设置。声景增强系统还可以包括播放系统,播放系统被设置为播放或再现一个或多个最佳掩蔽设置。



1. 声景增强系统,包括:

数据获取系统,所述数据获取系统被设置为提供周围声景数据;

数据库,所述数据库包括多个掩蔽设置;

感知属性预测器,所述感知属性预测器耦接于所述数据获取系统和所述数据库,所述感知属性预测器被设置为基于所述周围声景数据,为所述多个掩蔽设置中的每个掩蔽设置,生成在一个或多个预定义感知属性指标上代表感知的预测;

掩蔽设置排序系统,所述掩蔽设置排序系统被设置为基于所述感知属性预测器生成的所述预测,确定一个或多个最佳掩蔽设置;以及

播放系统,所述播放系统被设置为播放或再现所述一个或多个最佳掩蔽设置。

2. 根据权利要求1所述的声景增强系统,其中所述数据获取系统还被设置为提供群体统计或情境数据。

3. 根据权利要求2所述的声景增强系统,其中所述感知属性预测器被设置为还基于所述群体统计或情境数据,生成所述预测。

4. 根据权利要求2至3中任一项所述的声景增强系统,其中所述群体统计或情境数据包括与环境参数相关的数据、一名或多名听众的群体统计数据、所述一名或多名听众的心理数据、所述一名或多名听众的生理数据或其任意组合。

5. 根据权利要求4所述的声景增强系统,其中所述与环境参数相关的数据包括与位置相关的数据、与所述视觉环境相关的数据、附近的人数、空气温度、湿度、风速或其任意组合。

6. 根据权利要求4或5所述的声景增强系统,其中所述一名或多名听众的所述群体统计数据包括与年龄、性别、职业或其任意组合相关的数据。

7. 根据权利要求4至6中任一项所述的声景增强系统,其中所述一名或多名听众的所述心理数据包括与噪声敏感度、感知压力、幸福感指数评分或其任意组合相关的数据。

8. 根据权利要求4至7中任一项所述的声景增强系统,其中所述一名或多名听众的所述生理数据包括与心率、血压、体温或其任意组合相关的数据。

9. 根据权利要求1至8中任一项所述的声景增强系统,其中所述感知属性预测器被设置为基于一个或多个掩蔽增益输入,生成预测。

10. 根据权利要求1至9中任一项所述的声景增强系统,其中所述一个或多个预定义感知属性指标包括愉悦度、活跃度、事件性、平静度、感知响度、声音质量、锐度、粗糙度或其任意组合。

11. 根据权利要求1至10中任一项所述的声景增强系统,其中所述感知属性预测器被设置为通过组合所述周围声景数据和每个掩蔽设置,以生成所述预测。

12. 根据权利要求1至11中任一项所述的声景增强系统,

其中所述感知属性预测器包括设置为提取所述周围声景数据的特征的声景特征提取器;

其中所述感知属性预测器包括设置为提取每个掩蔽设置的特征的掩蔽特征提取器;并且

其中所述感知属性预测器被设置为基于所述周围声景数据的提取特征和每个掩蔽设置的提取特征,生成所述预测。

13. 根据权利要求1至12中任一项所述的声景增强系统,

其中所述感知属性预测器包括设置为从所述周围声景数据和每个掩蔽设置提取客观特征的客观特征块;

其中所述感知属性预测器包括设置为从所述群体统计或情境数据提取主观特征的主观特征块;并且

其中所述感知属性预测器被设置为基于所述提取的客观特征和所述提取的主观特征,生成所述预测。

14. 根据权利要求1至13中任一项所述的声景增强系统,其中所述感知属性预测器包括线性回归模型,所述模型线性回归模型被设置为基于声学或心理声学参数,生成所述预测,所述声学或心理声学参数是基于所述周围声景数据和每个掩蔽设置计算的。

15. 根据权利要求1至14中任一项所述的声景增强系统,其中所述感知属性预测器包括一个或多个深度神经网络,所述一个或多个深度神经网络被设置为基于原始音频、基于所述周围声景数据和每个掩蔽设置计算的频谱图呈现,或者所述原始音频和所述频谱图呈现的组合,生成所述预测。

16. 根据权利要求1至15中任一项所述的声景增强系统,其中所述数据获取系统被设置为基于一个或多个收音器、一根或多根天线、存储介质或其任意组合的输入,生成所述周围声景数据。

17. 根据权利要求1至16中任一项所述的声景增强系统,其中所述多个掩蔽设置包括一个或多个录制的或合成的声音的音轨、一个或多个静音的音轨,或者从所述一个或多个录制的或合成的声音的音轨和所述一个或多个静音的音轨衍生出的一个或多个音轨。

18. 根据权利要求1至17中任一项所述的声景增强系统,其中所述播放系统包括一个或多个扬声器、一个或多个虚拟现实或增强现实耳麦、一个或多个耳机、一个或多个耳塞或其任意组合。

19. 根据权利要求1至18中任一项所述的声景增强系统,其中生成的所述预测是不确定性的。

20. 一种声景增强系统的形成方法,所述方法包括:

提供数据获取系统,所述数据获取系统设置为提供周围声景数据;

提供数据库,所述数据库包括多个掩蔽设置;

将感知属性预测器耦接于所述数据获取系统和所述数据库,所述感知属性预测器被设置为基于所述周围声景数据,为所述多个掩蔽设置中的每个掩蔽设置生成在一个或多个预定义感知属性指标上代表感知的预测;

提供掩蔽设置排序系统,所述掩蔽设置排序系统被设置为基于所述感知属性预测器生成的所述预测,确定一个或多个最佳掩蔽设置;以及

提供播放系统,所述播放系统设置为播放或再现所述一个或多个最佳掩蔽设置。

声景增强系统及其形成方法

[0001] 相关申请的交叉引用

[0002] 本申请要求于2022年4月27日提交的新加坡申请号10202204451S的优先权的权益,其内容在此通过援引整体并入本文,以用于所有目的。

技术领域

[0003] 本申请的实施例涉及声景增强系统。本申请的实施例涉及声景增强系统的形成方法。

背景技术

[0004] 世界卫生组织(World Health Organization,WHO)已经将噪声暴露列为排在空气污染之后的第二大环境污染。噪声暴露也被比作二手烟。采取行动以减轻噪声暴露的迫切呼吁源于噪声对健康的影响的、越来越多的证据,噪声对健康的影响是例如缺血性心脏病发病率的上升风险、烦躁和睡眠障碍。世界卫生组织还强调环境噪声暴露对心理健康和幸福感(well-being)的影响,其中新证据表明对抑郁和焦虑指标的不利影响。

[0005] 从噪声管理到声景管理的范式转变(paradigm shift)正在出现。ISO12913-1将“声景”定义为“所感知的或所体验的声学环境和/或被一个人或多个人在情境中所理解的声学环境”。与噪声管理相反,声景管理框架将声音视为资源而不是废物;声景管理框架专注于偏爱的声音而不是令人不适的声音;并且声景管理框架以期望出现的声音进行管理,以掩蔽不期望出现的声音,并且管理以减少不期望出现的声音,以代替仅降低声级(sound levels)。因此,有基于“掩蔽”声音的增强或在声学环境中引入“掩蔽”声音的声景干预技术,以改善声学舒适度的整体感知。通常,该干预涉及自然声音的增强(例如通过扬声器),并且已在户外休闲场所和例如疗养院的室内进行了试验,以提高声学舒适度。利用期望出现的声音的声景增强与声音掩蔽系统相似,声音掩蔽系统通常在办公室环境中使用,以减少干扰。声音掩蔽系统与基于声景的增强系统之间的关键区别在于掩蔽是基于客观指标,然而声景增强则主要基于以情境为基础的主观指标。情境“包括人与活动和地点之间的、在空间和时间上的相互关系”,并且在ISO12913-1中通过示例进一步详细说明。还值得注意的是,声景不仅限于现实世界的环境,还可以包括虚拟环境,并且甚至可以包括记忆中的回忆。

发明内容

[0006] 本申请的实施例可以提供声景增强系统。声景增强系统可以包括数据获取系统,数据获取系统被设置为提供周围声景数据。声景增强系统还可以包括数据库,数据库包括多个掩蔽设置。声景增强系统还可以包括感知属性预测器,感知属性预测器耦接于数据获取系统和数据库,感知属性预测器被设置为基于周围声景数据,为多个掩蔽设置中的每个掩蔽设置,生成在一个或多个预定义感知属性指标上代表感知的预测。声景增强系统可以附加地包括掩蔽设置排序系统,掩蔽设置排序系统被设置为基于感知属性预测器生成的预

测,确定一个或多个最佳掩蔽设置。声景增强系统还可以包括播放系统,播放系统被设置为播放或再现一个或多个最佳掩蔽设置。

[0007] 本申请的实施例可以涉及声景增强系统的形成方法。该方法可以包括提供数据获取系统,数据获取系统被设置为提供周围声景数据。该方法还可以包括提供数据库,数据库包括多个掩蔽设置。该方法可以进一步包括将感知属性预测器耦接于数据获取系统和数据库,感知属性预测器被设置为基于周围声景数据,为多个掩蔽设置中的每个掩蔽设置,生成在一个或多个预定义感知属性指标上代表感知的预测。该方法可以附加地包括提供掩蔽设置排序系统,掩蔽设置排序系统被设置为基于感知属性预测器生成的预测,确定一个或多个最佳掩蔽设置。该方法可以进一步包括提供播放系统,播放系统被设置为播放或再现一个或多个最佳掩蔽设置。

附图说明

[0008] 在附图中,相似的参考符号在不同视图中通常指相同的部件。附图不一定按比率绘制,而是实质上着重于说明各个实施例的原理。在以下描述中,将参考以下附图描述本发明的各个实施例。

[0009] 图1示出根据各个实施例的声景增强系统的示意图。

[0010] 图2示出根据各个实施例的、声景增强系统的形成方法的示意图。

[0011] 图3示出根据各个实施例的声景增强系统的示意图。

[0012] 图4示出ISO12913-3提出的情感质量属性的二维圆周八分模型,根据各个实施例的感知属性预测器的预测可以基于该二维圆周八分模型。

[0013] 图5是根据各个实施例的感知属性预测器的示意图。

[0014] 图6是根据各个实施例的感知属性预测器的示意图。

[0015] 图7是根据各个实施例的感知属性预测器的示意图。

[0016] 图8示出根据各个实施例的自动掩蔽选择系统(automatic masker selection system,AMSS)的训练和推理方案。

[0017] 图9示出(a)根据各个实施例使用的基本卷积循环神经网络(convolutional recurrent neural network,CRNN)架构;(b)根据各个实施例的、(a)中的特征映射块的、一个可能的实施;以及(c)根据各个其他实施例的、(a)中的特征映射块的、其他可能的实施。

[0018] 图10是示出根据各个实施例的、对每个设置测试的10次运行(上:交叉验证集,下:测试集)中的概率感知属性预测器(probabilistic perceptual attribute predictor,PPAP)的均折(mean fold)均方误差(mean squared errors,MSEs)($\pm$ 标准差)的表格。

[0019] 图11示出使用测试集上的50个模型(点积变量)进行的掩蔽选择,其中具有根据各个实施例的、朴素最大 μ_k 值选择方案(左)以及根据各个实施例的随机采样方案,以激励掩蔽器探索(右)。

[0020] 图12A示出根据各个实施例的概率感知属性预测器(probabilistic perceptual attribute predictor,PPAP)的示意图。

[0021] 图12B示出各个实施例中根据计算式(8)、(9)和(10)的三种特征增强方法。

[0022] 图13示出根据各个实施例的、减少总运行时间的算法。

[0023] 图14示出(上图)均方误差(mean squared error,MSE)作为注意力块类型的函数

的图示,该图示出根据各个实施例的、每个注意力块类型和特征增强方法的愉悦度预测的验证MSE的小提琴图示;以及(下图)平均绝对误差(mean absolute error,MAE)作为注意力块类型和特征增强方法的函数的图示,该图示出根据各个实施例的、愉悦度预测的验证MAE的小提琴图示。

[0024] 图15A是根据各个实施例的、ISO愉悦度作为对数增益(log gain)的函数的图示,该图示出使用加性注意力(additive attention,AA)和CONV增强(种子5、验证折2)的、为掩蔽器类别“水”进行的增益插值(gain interpolation)。

[0025] 图15B是根据各个实施例的、ISO愉悦度作为对数增益的函数的图示,该图示出使用具有加性注意力(AA)和CONV增强(种子5,验证折2)的、为掩蔽器类别“交通”进行的增益插值。

[0026] 图15C是根据各个实施例的、ISO愉悦度作为对数增益的函数的图示,该图示出使用具有加性注意力(AA)和CONV增强(种子5、验证折2)的、为掩蔽器类别“鸟”进行的增益插值。

[0027] 图15D是根据各个实施例的、ISO愉悦度作为对数增益的函数的图示,该图示出使用加性注意力(AA)和CONV增强(种子5、验证折2)的、为掩蔽器类别“施工”进行的增益插值。

[0028] 图15E是根据各个实施例的、ISO愉悦度作为对数增益的函数的图示,该图示出使用加性注意力(AA)和CONV增强(种子5、验证折2)的、为掩蔽器类别“静音”进行的增益插值。

具体实施方式

[0029] 以下详细描述参考附图,附图以说明的方式示出可以实施本发明的具体细节和实施例。这些实施例被足够详细地描述,以使本领域技术人员能够实施本发明。可以使用其他实施例,并且可以在不脱离本发明范围的情况下,进行结构方面、逻辑方面和电方面的改变。各个实施例不一定是相互排斥的,因为一些实施例可以与一个或多个其他实施例组合以形成新的实施例。

[0030] 在一个实施例的上下文中描述的特征可以相应地适用于其他实施例中相同或相似的特征。在一个实施例的上下文中描述的特征可以相应地适用于其他实施例,即使在这些其他实施例中没有明确描述。此外,在一个实施例的上下文中针对一个特征描述的添加和/或组合和/或替代可以相应地适用于其他实施例中相同或相似的特征。

[0031] 在各个实施例的上下文中,关于特征或元件使用的冠词“一”、“一个”和“该”包括对一个或多个特征或元件的引用。

[0032] 在各个实施例的上下文中,应用于数值的术语“约”或“大约”涵盖确切值和合理差值,合理差值在例如指定值的10%以内。

[0033] 如本文所用,术语“和/或”包括所列相关项目中的一个或多个项目的任意组合和全部组合。

[0034] “包括”指包括但不限于“包括”一词后面的任何内容。因此,术语“包括”的使用表示所列元件是必需的或强制性的,但是其他元件是可选的,并且其他元件可以存在或不存在。

[0035] “由……组成”指包括且限于短语“由……组成”之间的任何内容。因此,短语“由……组成”表示所列元件是必需的或强制性的,并且不存在其他元件。

[0036] 针对一个声景增强系统的上下文中描述的实施例对于其他声景增强系统同样有效。类似地,在方法的上下文中描述的实施例对于声景增强系统同样有效,反之亦然。

[0037] 目前最先进的增强或掩蔽干预涉及通过扬声器的声音静态播放,其中必须手动选择声级和音轨。各个实施例可以利用首创的人工智能模型,以自动选择音轨,该音轨在最合适的声级下产生任意主观指标(例如感知响度、ISO12913-3的圆周八分指标(Circumplex Octant Scale)-愉悦度(pleasantness)、烦扰度(annoyance)、宁静度(tranquillity)、平静度(calmness)、活跃度(vibrancy)、事件性(eventfulness)等)上的最高(或最低)值。人工智能(artificial intelligence, AI)模型是由南洋理工大学(Nanyang Technological University, NTU)研究团队在基于ISO12913-2标准收集的新加坡当地人群的人类主观反应的大数据集上进行训练的。ISO12913系列标准代表声音环境管理的范式转变,并且详细描述基于感知的方法以全面评估和分析声音环境。此外,人工智能模型还可以在不按照ISO12913-2的要求进行繁琐且耗力的调查或问卷的情况下,兼作自动评估声音环境感知质量的预测工具。

[0038] 图1是根据各个实施例的声景增强系统的示意图。声景增强系统可以包括数据获取系统102,数据获取系统102被设置为提供周围声景数据。声景增强系统还可以包括数据库104,数据库104包括多个掩蔽设置。声景增强系统可以进一步包括感知属性预测器106,感知属性预测器106耦接于数据获取系统102和数据库104,感知属性预测器106被设置为基于周围声景数据,为多个掩蔽设置中的每个掩蔽设置,生成在一个或多个预定义感知属性指标上代表感知的预测。声景增强系统可以附加地包括掩蔽设置排序系统108,掩蔽设置排序系统108被设置为基于感知属性预测器106生成的预测,确定一个或多个最佳掩蔽设置。声景增强系统还可以包括播放系统110,播放系统110被设置为播放或再现一个或多个最佳掩蔽设置。

[0039] 换言之,声景增强系统可以包括感知属性预测器106,感知属性预测器106可以连接于数据获取系统102和数据库104。声景增强系统还可以包括连接于感知属性预测器106的掩蔽设置排序系统108和连接于掩蔽设置排序系统108的播放系统110。

[0040] 为避免疑问,图1旨在说明根据各个实施例的声景增强系统的一些特征,并且不旨在限制例如各个组件的尺寸、形状、取向、设置等。

[0041] 在各个实施例中,数据获取系统102还可以被设置为提供群体统计或情境数据。感知属性预测器106还可以被设置为基于群体统计或情境数据生成预测。群体统计或情境数据可以由数据获取系统102直接接收、测量或感测,或者从外部源接收,外部源是例如便携式健康监测设备或存储设备等设备,如下文更详细描述。

[0042] 群体统计或情境数据可以包括与环境参数相关的数据、一名或多名听众的群体统计数据、一名或多名听众的心理数据、一名或多名听众的生理数据或其任意组合。

[0043] 就此而言,一名或多名听众可以包括真实的听众和/或假想的听众。

[0044] 在各个实施例中,与环境参数相关的环境数据可以是或者可以包括与位置相关的数据、与视觉环境相关的数据、附近的人数、空气温度、湿度、风速或其任意组合。环境参数可以是非声学环境参数。例如,数据获取系统102可以包括或者可以耦接于全球定位系统(Global Positioning System, GPS)传感器,以确定与位置相关的数据。数据获取系统102可以包括或者可以耦接于摄像头(camera),以确定与附近人数相关的数据和/或从环境中

获得视觉信息(例如,确定特定类型的家具、宠物等的存在,确定人的移动,确定存在的绿化物或水的百分比等)。数据获取系统102可以包括或者可以耦接于温度计,以确定与空气温度相关的数据。数据获取系统102可以包括或者可以耦接于湿度计,以确定与湿度相关的数据。数据获取系统102可以包括或者可以耦接于风速计,以确定与风速相关的数据。在该上下文中,术语“耦接”可以指直接连接或间接连接。连接可以通过例如有线方式或无线方式建立。

[0045] 在各个实施例中,一名或多名听众的群体统计数据可以是或者可以包括与年龄、性别、职业或其任意组合相关的数据。在各个实施例中,数据获取系统102提供的一名或多名听众的群体统计数据可以由数据获取系统102从外部源获得。在各个实施例中,一名或多名听众的群体统计数据可以通过附连于数据获取系统102的数字化形式提供,或者可以作为耦接于(例如,直接耦接或者通过基于云的服务器耦接)数据获取系统102的存储设备上的预加载数据提供。在各个实施例中,一名或多名听众的群体统计数据可以手动提供给数据获取系统102,例如手动键入数据获取系统102,或者根据用户提供的手动指令从远程系统发送至数据获取系统102。在各个实施例中,数据获取系统102可以被设置为自动获得一名或多名听众的群体统计数据,例如,通过有线方式或无线方式从远程系统获得。

[0046] 在各个实施例中,一名或多名听众的心理数据可以是或者可以包括与噪声敏感度(使用例如韦恩斯坦噪声敏感度量表(Weinstein Noise Sensitivity Scale))、感知压力(使用例如科恩感知压力量表(Cohen's Perceived Stress Scale))、幸福感指数得分(使用例如WHO-5幸福感指数(WHO-5 Well-Being Index))或其任意组合相关的数据。在各个实施例中,数据获取系统102提供的、一名或多名听众的心理数据可以由数据获取系统102从外部源获得。在各个实施例中,一名或多名听众的心理数据可以通过附连于数据获取系统102的数字化形式提供,或者可以作为耦接于(例如,直接耦接或通过基于云的服务器耦接)数据获取系统102的存储设备上的预加载数据提供。在各个实施例中,一名或多名听众的心理数据可以手动提供给数据获取系统102,例如手动键入数据获取系统102,或者根据用户提供的手动指令从远程系统发送至数据获取系统102。在各个实施例中,数据获取系统102可以被设置为自动获得一名或多名听众的心理数据,例如,通过有线方式或无线方式从远程系统获得。

[0047] 在各个实施例中,一名或多名听众的生理数据可以是或者可以包括与心率、血压、体温或其任意组合相关的数据。例如,数据获取系统102可以包括或者可以耦接于(例如直接耦接或通过基于云的服务器耦接)一个或多个便携式健康监测设备,该一个或多个便携式健康监测设备被设置为确定或测量一名或多名听众的生理数据。数据获取系统102可以被设置为通过有线方式或无线方式,从一个或多个便携式健康监测设备获得一名或多名听众的生理数据。

[0048] 在各个实施例中,感知属性预测器106可以被设置为还基于一个或多个掩蔽增益输入以生成预测。例如,一个或多个掩蔽增益输入可以是一个或多个数字增益级别或一个或多个掩蔽增益波形。

[0049] 在各个实施例中,一个或多个预定义感知属性指标可以包括愉悦度、活跃度、平静度、感知响度(perceived loudness)、声音质量(sound quality)、锐度(sharpness)、粗糙度(roughness)或其任意组合。在各个实施例中,一个或多个预定义感知属性指标可以从

ISO12913-3中定义的情感质量属性中选择。

[0050] 在各个实施例中,生成的预测可以是或者可以不是单个数值。在各个实施例中,生成的预测可以是非确定性的。在各个其他实施例中,生成的预测可以是确定性的。在各个实施例中,预测可以是概率分布、从概率分布中提取的随机值、一个或多个预定义感知属性的向量或多变量呈现。在各个实施例中,生成的预测可以是预测的属性分布。

[0051] 在各个实施例中,感知属性预测器106可以是或者可以包括概率感知属性预测器。在各个实施例中,感知属性预测器106可以被设置为通过混合、组合或添加周围声景数据和每个掩蔽设置,以生成预测。在各个实施例中,周围声景数据和每个掩蔽设置可以被添加、混合或组合于例如推理引擎(inference engine)中,以在将增强声景数据输入感知属性预测器106之前形成增强声景数据。每个掩蔽设置的掩蔽波形可以与例如掩蔽增益波形的掩蔽增益输入加权(即相乘),以生成加权掩蔽波形。加权掩蔽波形可以与周围声景数据的周围声景波形组合(即添加于周围声景数据的周围声景波形),以生成增强声景波形。感知属性预测器106的预测块可以接收增强声景波形,以生成预测。

[0052] 在各个实施例中,感知属性预测器106可以包括设置为提取周围声景数据的特征的声景特征提取器。感知属性预测器106可以包括设置为提取每个掩蔽设置的特征的掩蔽特征提取器。感知属性预测器106可以包括特征级增强器,特征级增强器被设置为基于周围声景数据的提取特征、每个掩蔽设置的提取特征和例如数字增益级别的掩蔽增益输入,生成一个或多个增强声景特征。感知属性预测器106可以包括预测块,预测块被设置为基于一个或多个增强声景特征、周围声景数据的提取特征和每个掩蔽设置的提取特征,生成预测。感知属性预测器106可以被设置为基于周围声景数据的提取特征和每个掩蔽设置的提取特征,生成预测。

[0053] 在各个实施例中,感知属性预测器106可以包括客观特征块,客观特征块被设置为从周围声景数据和每个掩蔽设置提取客观特征。感知属性预测器106可以被进一步设置为还基于掩蔽增益输入,提取客观特征。感知属性预测器106可以包括主观特征块,主观特征块被设置为从群体统计或情境数据中提取主观特征。主观特征块还可以被设置为处理提取的主观特征和提取的客观特征,以生成预测。感知属性预测器106可以被设置为基于提取的客观特征和提取的主观特征,生成预测。

[0054] 在各个实施例中,感知属性预测器106可以包括线性回归模型,该线性回归模型被设置为基于声学或心理声学参数,生成预测,该声学或心理声学参数是基于周围声景数据和每个掩蔽设置计算的。

[0055] 在各个实施例中,感知属性预测器106可以包括一个或多个深度神经网络,该一个或多个深度神经网络被设置为基于原始音频、频谱图呈现或原始音频和频谱图呈现的组合,生成预测,该频谱图呈现是基于周围声景数据和每个掩蔽设置计算的。

[0056] 在各个实施例中,感知属性预测器106可以包括如本文所述的复合网络或如本文所述的网络和/或模型的任意组合。

[0057] 在各个实施例中,周围声景数据可以由数据获取系统102直接接收、记录、测量或感测,或者可以从外部源接收。在各个实施例中,数据获取系统102可以被设置为基于记录设备、接收设备和/或存储设备的输入,提供或生成周围声景数据。数据获取系统102可以被设置为基于一个或多个收音器(microphones)、一根或多根天线、存储介质或其任意组合的

输入,提供或生成周围声景数据。在各个实施例中,数据获取系统102可以被设置为通过直接记录或感测环境的周围声景,以提供或生成周围声景数据。

[0058] 在各个实施例中,多个掩蔽设置可以包括一个或多个录制的或合成的声音的音轨、一个或多个静音的音轨或者从该一个或多个录制的或合成的声音的音轨和该一个或多个静音的音轨衍生出的一个或多个音轨。

[0059] 在各个实施例中,播放系统110可以包括一个或多个扬声器、一个或多个虚拟现实或增强现实耳麦、一个或多个耳机、一个或多个耳塞或其任意组合。

[0060] 在各个实施例中,可以使用一个或多个计算、处理或电子设备或系统,以实施声景增强系统的各个组件。在一个示例中,数据获取系统102、数据库104、感知属性预测器106、掩蔽设置排序系统108、播放系统110可以在包括收音器和扬声器的单个计算或处理设备或系统中实施,该单个计算或处理设备或系统是例如物理服务器或基于云的服务器、个人计算机系统或移动设备。在另一示例中,数据获取系统102可以通过(具有收音器的)移动设备实施,与此同时,数据库104、感知属性预测器106、掩蔽设置排序系统108和播放系统110可以通过个人计算机系统实施,该个人计算机系统具有直接地或间接地与移动设备无线通信(例如WiFi或蓝牙)的扬声器。在另一示例中,数据获取系统102可以用远程连接于收音器和环境传感器的笔记本电脑实施。数据库104、感知属性预测器106和掩蔽设置排序系统108可以使用通过服务器与笔记本电脑通信的另一计算设备实施。播放系统110可以是与计算机设备无线通信的耳机。

[0061] 图2示出根据各个实施例的声景增强系统的形成方法的示意图。该方法可以包括在步骤202提供数据获取系统,该数据获取系统被设置为提供周围声景数据的。该方法还可以包括在步骤204提供包括多个掩蔽设置的数据库。该方法可以进一步包括在步骤206将感知属性预测器耦接于数据获取系统和数据库,该感知属性预测器被设置为基于周围声景数据,为多个掩蔽设置中的每个掩蔽设置,生成在一个或多个预定义感知属性指标上代表感知的预测。该方法可以附加地包括在步骤208提供掩蔽设置排序系统,该掩蔽设置排序系统被设置为基于感知属性预测器生成的预测,确定一个或多个最佳掩蔽设置。该方法可以进一步包括在步骤210提供播放系统,该播放系统被设置为播放或再现一个或多个最佳掩蔽设置。

[0062] 换言之,该方法可以包括将感知属性预测器耦接于数据获取系统和包含多个掩蔽设置的数据库。该方法还可以包括将掩蔽设置排序系统耦接于感知属性预测器,并且将播放系统耦接于掩蔽设置排序系统。

[0063] 为避免疑问,图2不旨在限制各个步骤的顺序。例如,步骤202可以在步骤204之前、之后或同时发生。

[0064] 在各个实施例中,数据获取系统还可以被设置为提供群体统计或情境数据。感知属性预测器还可以被设置为基于群体统计或情境数据,生成预测。

[0065] 在各个实施例中,群体统计或情境数据可以包括与环境参数相关的数据、一名或多名听众的群体统计数据、一名或多名听众的心理数据、一名或多名听众的生理数据或其任意组合。

[0066] 在各个实施例中,与环境参数相关的数据可以包括与位置相关的数据、与视觉环境相关的数据、附近的人数、空气温度、湿度、风速或其任意组合。

[0067] 在各个实施例中,一名或多名听众的群体统计数据可以包括与年龄、性别、职业或其任意组合相关的数据。

[0068] 在各个实施例中,一名或多名听众的心理数据可以包括与噪声敏感度、感知压力、幸福感指数得分或其任意组合相关的数据。

[0069] 在各个实施例中,一名或多名听众的生理数据可以包括与心率、血压、体温或其任意组合相关的数据。

[0070] 在各个实施例中,感知属性预测器可以被设置为基于一个或多个掩蔽增益输入,生成预测。

[0071] 在各个实施例中,一个或多个预定义感知属性指标可以包括愉悦度、活跃度、事件性、平静度、感知响度、声音质量、锐度、粗糙度或其任意组合。

[0072] 在各个实施例中,感知属性预测器可以被设置为通过组合周围声景数据和每个掩蔽设置,以生成预测。

[0073] 在各个实施例中,感知属性预测器可以包括声景特征提取器,该声景特征提取器被设置为提取周围声景数据的特征。感知属性预测器可以包括掩蔽特征提取器,该掩蔽特征提取器被设置为提取每个掩蔽设置的特征。感知属性预测器可以被设置为基于周围声景数据的提取特征和每个掩蔽设置的提取特征,生成预测。

[0074] 在各个实施例中,感知属性预测器可以包括客观特征块,该客观特征块被设置为从周围声景数据和每个掩蔽设置中提取客观特征。感知属性预测器可以包括主观特征块,该主观特征块被设置为从群体统计或情境数据中提取主观特征。感知属性预测器可以被设置为基于提取的客观特征和提取的主观特征,生成预测。

[0075] 在各个实施例中,感知属性预测器可以包括线性回归模型,该线性回归模型被设置为基于声学或心理声学参数生成预测,该声学或心理声学参数是基于周围声景数据和每个掩蔽设置计算的。

[0076] 在各个实施例中,感知属性预测器可以包括一个或多个深度神经网络,该一个或多个深度神经网络被设置为基于原始音频、频谱图呈现、或原始音频和频谱图呈现的组合生成预测,频谱图呈现是基于周围声景数据和每个掩蔽设置计算的。

[0077] 在各个实施例中,数据获取系统可以被设置为基于一个或多个收音器、一根或多根天线、存储介质或其任意组合的输入,提供或生成周围声景数据。

[0078] 在各个实施例中,多个掩蔽设置可以包括一个或多个录制的或合成的声音的音轨、一个或多个静音的音轨或从该一个或多个录制的或合成的声音的音轨和该一个或多个静音的音轨衍生出的一个或多个音轨。

[0079] 在各个实施例中,播放系统可以包括一个或多个扬声器、一个或多个虚拟现实或增强现实耳麦、一个或多个耳机、一个或多个耳塞或其任意组合。

[0080] 在各个实施例中,生成的预测可以是非确定性的(non-deterministic)。

[0081] 图3示出根据各个实施例的声景增强系统的示意图。声景增强系统也可以称为自动掩蔽选择系统(automatic masker selection system,AMSS)。声景增强系统可以包括数据获取系统302,数据获取系统302负责提供周围声景数据312,即声学环境的音频和/或直接从声学环境的音频中计算的参数。为从声学环境获得周围声景数据312,数据获取系统302可以使用例如一个或多个收音器的实时记录、从无线天线接收的数字音频信号、存储于

附连的存储介质上的数据或其任意组合。

[0082] 此外, 声景增强系统可以包括数据库, 该数据库包括多个掩蔽设置304 (替代地称为候选掩蔽设置组)。术语“掩蔽设置”可以指可通过任何形式的播放添加于312中的声学环境的、与音轨的预期播放所需的其他参数或效果相结合的任何音轨。掩蔽设置可以替代地称为“掩蔽器”。该参数或效果可以包括例如数字增益级别、空间化 (spatialization)、动态包络 (dynamic envelope) 和/或滤波 (filtering) 的应用。掩蔽设置的示例可以包括但不限于由以不同音量播放的录制的或合成的声音组成的音轨、仅由静音组成的音轨和/或从304中的一个或多个其他音轨的函数中衍生的音轨, 该仅由静音组成的音轨的播放不会对312中的声学环境构成任何声音添加。掩蔽设置可以替代地或附加地指选择掩蔽器的过程, 例如从音轨的数据库304中选择掩蔽器的过程。掩蔽设置策略可以是自主的 (autonomous), 并且可以 (通过例如监控收音器或通过数字化输入) 适应当前声景的动态 (例如振幅、频率内容、情境)。

[0083] 数据获取系统302还可以可选地负责采集群体统计或情境数据314作为声学环境本身的辅助参数, 并且负责提供群体统计或情境数据314。该群体统计或情境数据314可以包括但不限于 (1) 非声学环境参数, 例如位置 (通过GPS传感器采集)、附近人数 (通过摄像头采集)、空气温度 (通过温度计采集)、湿度 (通过湿度计采集) 和/或风速 (通过风速计采集); (2) 声学环境中真实听众或假想听众的群体统计数据 (例如, 通过附连于系统的数字化形式 (digital form) 采集或者作为存储设备上的预加载数据), 例如年龄、性别和/或职业; (3) 声学环境中真实听众或假想听众的心理数据 (通过例如与群体统计方法相似的方法采集), 例如噪声敏感度 (使用例如韦恩斯坦噪声敏感度量表获得)、感知压力 (使用例如科恩感知压力量表获得), 以及/或者幸福感指数得分 (使用例如WHO-5幸福感指数获得); 和/或声学环境中真实听众或假想听众的生理数据 (通过例如便携式健康监测设备获得), 例如心率、血压和/或体温。

[0084] 数据获取系统302可以将周围声景数据312和可选的群体统计或情境数据314提供给感知属性预测器306。在给定周围声景数据312、数据库304的掩蔽设置以及可选的群体统计或情境数据314作为输入的情况下, 感知属性预测器306可以为数据库304中的每个掩蔽设置, 输出在一个或多个预定义感知属性指标上的预测 (对314中的数据所对应的听众的预测), 如同每个掩蔽设置的播放将在声学环境中实现。例如, 一个或多个预定义感知属性指标可以源自用于数据库304中的每个掩蔽设置的ISO12913-3的圆周八分指标。例如, 一个或多个预定义感知属性指标可以基于用于数据库304中每个掩蔽设置的愉悦度、事件性、声音质量、感知响度、锐度、粗糙度等。图4示出ISO12913-3中呈现的情感质量属性的二维圆周八分模型, 根据各个实施例的感知属性预测器306的预测可以基于该二维圆周八分模型。

[0085] 在图3中, 该预测的集合由316表示。预测316可以是或者可以不是单个数值。类似地, 在各个实施例中, 生成的预测316可以是非确定性的, 与此同时, 在各个其他实施例中, 生成的预测316可以是确定性的。术语“确定性的”可以指, 每当周围声景数据312、数据库304的多个掩蔽设置以及可选的群体统计或情境数据314的相同输入被提供给感知属性预测器306时, 就会产生相同的输出。概率分布 (或从概率分布中抽取的随机值)、向量或感知属性的任何其他多变量呈现可以是作为预测316的可能输出。

[0086] 感知属性预测器306可以是或者可以包括从周围声景数据312、数据库304的多个

掩蔽设置以及可选的群体统计或情境数据314中获取相同的输入以输出预测316的任何模型。该模型可以在将预测316作为输出输送之前,内部转换周围声景数据312、数据库304的多个掩蔽设置以及群体统计或情境数据314。该模型的示例可以包括但不限于:(1)线性回归模型,该线性回归模型获取从周围声景数据312计算的声学或心理声学参数、数据库304的多个掩蔽设置以及从群体统计或情境数据314计算的聚合平均值,以输出预测316;(2)深度神经网络,该深度神经网络获取从周围声景数据312中计算的呈现、数据库304的多个掩蔽设置,其中串联有群体统计或情境数据314的呈现,以输出预测316;(3)如图5至图7所示的复合网络;和/或(4)如本文所述的任意组合。

[0087] 图5是根据各个实施例的感知属性预测器506的示意图。在各个实施例中,感知属性预测器506可以被设置为通过将周围声景数据和每个掩蔽设置组合,以生成预测,例如,在推理引擎中以预校准的声景与掩蔽器的比率(soundscapes-to-masker ratios)进行组合。每个掩蔽设置的掩蔽波形504a可以与例如掩蔽增益波形的掩蔽增益输入504b加权(即相乘),以生成加权掩蔽波形。加权掩蔽波形可以与周围声景数据的周围声景波形512组合(即添加于周围声景数据的周围声景波形512),以生成增强声景波形518。感知属性预测器506的预测块520可以接收增强声景波形518,以生成预测510。

[0088] 在将增强声景波形输入感知属性预测器中之前,各个其他实施例可以不需要混合、组合或添加周围声景数据和每个加权掩蔽波形以形成增强声景波形。在将增强声景波形输入感知属性预测器之前不需要通过混合、组合或添加周围声景波形和每个加权掩蔽波形以形成增强声景数据,各个实施例可以在实际部署中允许实现较高效的计算和带宽系统。

[0089] 图6是根据各个实施例的感知属性预测器606的示意图。在各个实施例中,感知属性预测器606可以包括声景特征提取器622,声景特征提取器622被设置为提取周围声景数据的特征624,即提取周围声景数据的周围声景波形612的特征624。感知属性预测器606可以包括掩蔽特征提取器626,掩蔽特征提取器626被设置为提取每个掩蔽设置的特征628,即提取每个掩蔽设置的掩蔽波形604a的特征628。感知属性预测器606可以包括特征级别增强器630,特征级别增强器630被设置为基于周围声景数据的提取特征624、每个掩蔽设置的提取特征628和例如数字增益级别的掩蔽增益输入604b,生成一个或多个增强声景特征632。感知属性预测器606可以包括预测块620,预测块620被设置为基于一个或多个增强声景特征632、周围声景数据的提取特征624和每个掩蔽设置的提取特征628,生成预测610。概括而言,感知属性预测器606可以被设置为基于周围声景数据的提取特征624和每个掩蔽设置的提取特征628,生成预测610。

[0090] 图7是根据各个实施例的感知属性预测器706的示意图。在各个实施例中,感知属性预测器706可以包括客观特征块734,客观特征块734被设置为从周围声景数据和每个掩蔽设置中提取客观特征736,即从周围声景数据的周围声景波形712和每个掩蔽设置的掩蔽波形704a中提取客观特征736。感知属性预测器706可以被进一步设置为还基于掩蔽增益输入704b,提取客观特征736。感知属性预测器706可以包括主观特征块738,主观特征块738被设置为从群体统计或情境数据714中提取主观特征。主观特征块738还可以被设置为处理提取的主观特征和提取的客观特征736,以生成预测710。感知属性预测器706可以被设置为基于提取的客观特征736和提取的主观特征,生成预测710。

[0091] 再参考图3,感知属性预测器306提供的预测316可以输入掩蔽设置排序系统308中,以生成单个最佳掩蔽设置340。掩蔽设置排序系统308可以通过任何合适的指标确定最佳性(optimality),该指标可以是但不限于感知响度、声音质量、锐度、粗糙度、图4所示的一个或多个情感质量属性和/或其任意组合中的最大值或最小值。替代地,可以设想感知属性预测器306提供的预测316可以输入掩蔽设置排序系统308中,以生成多个最佳掩蔽设置。对于给定的指标,可以有多个最佳掩蔽设置(例如,如果用户决定的指标中绑定有多个掩蔽设置)。

[0092] 最佳掩蔽设置340可以用作播放系统310的输入。播放系统310可以实现最佳掩蔽设置340在声学环境中的实际播放或再现。播放系统310可以包括但不限于一个或多个扬声器、一个或多个虚拟现实或增强现实耳麦、一个或多个耳机、一个或多个耳塞和/或其任意组合。

[0093] 各个实施例可以用于具有较差的声音质量的室外城市区域(例如高速公路旁的公园),以减轻周围声学环境中所感知的、噪声的有害影响,或者用于室内区域(例如疗养院内部),以改善周围声学环境的声学舒适度和整体质量。由于该系统有效地旨在改变周围声学环境的感知,拥有室内或室外区域的任何商业实体可能对该系统感兴趣,该室内或室外区域的声学环境需要保持于固定的或期望的条件下。

[0094] 各个实施例能够选择掩蔽器且将掩蔽器添加于现实的周围声景和假想的周围声景中。如果将任意音频输入数据获取系统以代替收音器的实时录音,则各个实施例还可以通过添加掩蔽器,用以创建抽象的“体验区”。任何有兴趣基于听众感知制作或修改虚拟现实、混合现实或增强现实中的声景的商业实体都会对这一点感兴趣。

[0095] 各个实施例可以涉及系统,该系统基于优化预测的感知响应(例如,感知响度、ISO12913-3的圆周八分指标——愉悦度、烦扰度、安宁度、平静度、活跃度、事件性等),将待引入现有声景的音轨(即掩蔽器)(通过例如扬声器播放或通过例如耳机进行数字化添加)选择且设置于混合/增强声景。

[0096] 各个实施例可以基于感知属性预测器对人类感知的预测,选择掩蔽设置或掩蔽器以添加于周围声学环境,该感知属性预测器根据周围声景数据(现实的或假想的周围声景数据)、候选掩蔽设置的数据以及可选的该周围声景的听众的群体统计或情境数据,输出感知属性指标的值或分布。这体现了“声景”的ISO12913的定义,即将声学环境视为由一个人或多个人在情境中所感知的声学环境。相比之下,现有方法可能基于周围声学环境的客观声学参数(例如频谱、声压级、声源类型和方向),选择掩蔽设置或掩蔽器,并且可能不考虑人类感知对掩蔽选择或掩蔽设置选择的主观影响。

[0097] 在各个实施例中,感知属性预测器可以近乎实时地输出感知预测(输入与预测之间的延迟小于10秒),并且因此可以自动实时地为随时间改变的声学环境,提出随时间改变的最佳掩蔽设置的建议。换言之,当声学环境随时间改变时,各个实施例可以自主地改变正在播放的掩蔽器(即音轨)和/或设置(例如增益级别、空间化)。相比之下,现有系统仅允许静态掩蔽设置(仅潜在地应用有动态音量/滤波器效果)或用户控制的掩蔽设置(即,一个人必须基于其自己的个人/专家对最优性的意见,手动改变设置)。

[0098] 在各个实施例中,用户可以将系统设置为在多个选择方案中自动选取最佳掩蔽器。选择方案可以是任意的,但是各个实施例可以允许常规解决方案所没有的以下选择方

案：(1) 概率top-k：对不同掩蔽器中的组合掩蔽器和现有声景的预测感知响应值进行排序，并且在给出前(或后)k个排序值的情况下，从k个掩蔽设置中随机选取一个掩蔽设置；(2) 概率分布：从系统生成的概率分布中抽取组合掩蔽器和现有声景的感知响应值的预测作为随机样本。

[0099] 现有系统和掩蔽选择策略可能不允许通过动态策略选择最佳掩蔽器，并且可能具有以下特征中的一个特征：(1) 手动选择：系统用户或管理员基于其自己的意见选择最佳掩蔽器；(2) 确定性：从不同掩蔽器中选取最佳掩蔽器作为组合的掩蔽器和现有声景的感知响应值的确定性的、预定义的函数。

[0100] 各个实施例可以包括一个或多个音频换能器(例如扬声器、耳机)和一个或多个传感器(例如收音器、环境传感器)，该一个或多个音频换能器和该一个或多个传感器相互耦接于一个或多个计算设备。各个实施例可以具有现有系统/解决方案所不具备的优势。

[0101] 例如，各个实施例可以采用非能量掩蔽。能量掩蔽利用人类听觉外周过程(peripheral human auditory processes)的物理限制。掩蔽声音的振幅和频率带宽被调整以使得目标噪声在内耳处是完全地或部分地无法被听到的(inaudible)。因此，通常采用例如声压级或心理声学参数的客观指标，以优化或选择能量掩蔽器。各个实施例可以参考(但不限于)ISO12913-1中定义的声景概念。在文献中，使用附加声音进行的声景增强通常也被称为掩蔽。掩蔽技术可以包括能量掩蔽和信息掩蔽(或感知掩蔽)，其中信息掩蔽由与声音的可察觉性的影响构成，该声音的可察觉性与较高级的大脑中心中的音频(audio in the higher brain centres)相关。因此，掩蔽器可以由感知因素(例如ISO12913-2中的感知情感属性、感知的烦扰度)决定，并且不是仅基于现有系统/解决方案中的客观指标。

[0102] 非能量掩蔽的优势可以在关于使用自然声音进行的、交通噪声的掩蔽的文献中得到证实，在该文献中，随着目标噪声级别的增加，能量掩蔽的感知响度的降低逐渐消失(nullified)。这与在较高级别的交通噪声下，提供信息掩蔽的掩蔽器以低最多6dB的级别呈现时，统计上显著的感知响度的降低和声景质量改善相反。此外，在使用鸟鸣声作为掩蔽器的情况下，也观察到相似的感知响度的显著降低和声景质量的改善，其中使用鸟鸣声作为掩蔽器无法在能量上掩蔽交通噪声，这进一步证明非能量掩蔽的有效性。

[0103] 通过辅助输入将音频输入情境化的能力可以允许各个实施例为任意目标声景(以代替仅限于现有文献中描述的办公环境)选择掩蔽器。根据各个实施例的掩蔽器也可以不限于从随机噪声生成的音轨，也不局限于现有文献中描述的自然环境的声音(natural sounds)。只要所期望出现的感知属性被优化(优化可以指“最大化”或“最小化”)，可以添加的掩蔽器的数量就可以不受限制。

[0104] 在各个实施例中，音频输入数据可以是不限格式的，并且可以是物理收音器的实时数据或从存储介质访问的音轨。类似地，根据各个实施例的掩蔽器可以流式传输(streamed)至扬声器(或耳机)或与存储介质的目标声景进行数字化混合，并且在虚拟环境中作为单个组合音轨播放。因此，各个实施例能够掩蔽现实世界环境以及虚拟环境中的声景(例如在元宇宙应用程序中)。

[0105] 现有文献中描述的动态掩蔽是基于客观指标的，例如声级、音频频谱，或者是通过与所识别的目标噪声源的相近度实现的。各个实施例可以基于预测的感知属性动态地或适应地调整掩蔽设置。掩蔽设置可以不限于声级或频谱调整。

[0106] 示例研究1

[0107] 介绍

[0108] 已知仅基于给定声学环境的声压级(sound pressure level,SPL)的指示通常不足以反映过度噪声造成的烦扰度级别和对生活质量的影响。这已经引起了噪声控制的声景方法的兴起,该方法着重于改善噪声感知属性的干预措施,以代替简单地降低SPL。关于声景的国际标准ISO12913系列旨在通过提供愉悦度和事件性的圆周模型以规范该方法,可以基于对周围声学环境的主观评估,根据该圆周模型比较干预措施。

[0109] 因此,许多研究已经利用声景增强技术,通过对城市或室内声学环境添加掩蔽器,以改变声景的感知,从而优化例如主观评价的愉悦度或平静度的指标。然而,掩蔽器的选择通常是任意的、专家指导的或是基于事后分析的。掩蔽器的任意选择可能在实现所期望出现的感知改变的方面是不可靠的,专家指导的选择是耗力且耗时的,并且事后分析可能无法泛化至在没有观察到的情境中的、没有看到的掩蔽器和声景。

[0110] 克服这些限制的方法是根据声景和掩蔽器的、声学上的多样化选择,训练预测模型以预测如被给予原始听觉刺激的人主观评估的、某个或某种感知属性的值。训练完成后,掩蔽器对没有看到的声景的模拟添加可以作为对模型的输入馈送,以获得所述感知属性的预测。然后,可以选择对属性产生最多提升或最多降低的掩蔽器作为最佳掩蔽器,以用于实时增强。

[0111] 各个实施例可以涉及用于自动掩蔽选择系统(automatic masker selection system,AMSS)的、通过优化声学环境的愉悦度以增强城市声景的神经学方法(neural approach)。通过使用概率输出方案,模型可以为每个声景与掩蔽器的组合预测愉悦度分布,以代替单个确定值,这允许明确获取预测的置信度以及预测的愉悦度。

[0112] 相关工作

[0113] 在声景研究中,开发用于感知属性的预测模型的研究主要集中于较简单的机器学习模型上,例如线性回归、支持向量机(support vector machines,SVM)和浅层多级感知器(shallow multilayer perceptrons)。机器学习模型使用基于声学测量、心理声学参数、环境特征或由主成分分析阐明的机器学习模型的某个或某种线性组合的输入特征。还提出了用于李克特量表(Likert scales)的标度指标(scaling metric),以解释用于开发这些系统的量表中的非线性。

[0114] 另一方面,最近的系统性综述示出,尽管深度神经网络在例如声音事件定位、检测和分类的较“客观”的任务中很普遍,但是利用深度神经网络预测声景感知属性值的研究却很少。就此而言,最重要的研究似乎是比较SVM、卷积神经网络(convolutional neural network,CNN)、长短期记忆网络(long short-term memory network,LSTM)和微调VGGish网络(fine-tuned VGGish network)在情绪-声景数据集(Emo-Soundscapes dataset)上的表现的研究。根据通过自我评估人偶模型(Self-Assessment Manikin)进行的主观配对比较,情绪-声景包含1213个按效价(愉悦度)和唤醒度(事件性)排序的音频片段。微调的VGGish模型和从零训练的CNN在预测音频片段的效价和唤醒度方面,分别呈现为具有最小的均方误差(mean squared errors,MSE)。最近较新的一项研究利用相似的模型,在使用共计约一千个音频片段的数据集的情况下,以与ISO12913相似的圆周模型,将音乐分类为四个象限中的一个象限,并且将音乐分类为中性、平静、快乐、悲伤、愤怒和恐惧的类别。然而,

声景通常包含的不仅仅是音乐,因此仍需研究其结果是否可以泛化于更广泛的声学刺激类别。

[0115] 此外,文献中的这些现有模型是确定性模型,即在不考虑用作基准真实值的主观评分中固有的不确定性的情况下,相同的声学环境总是映射于相同的预测值。然而,鉴于无法合理控制的因素,例如人的当前压力级别或噪声敏感度,个人给出的这些评分本质上是随机的。因此,各个实施例可以涉及概率方法,在该概率方法中,通过训练神经网络以预测输出的可能评分的分布,并且随后从该分布中摘取,这允许模型考虑每个独特声景中不同程度的不确定性。

[0116] 提出的方法

[0117] 图8示出根据各个实施例的自动掩蔽选择系统(automatic masker selection system,AMSS)的训练和推理示意图。

[0118] 考虑声景 S_0 和可以添加于声景的 K 个候选掩蔽器的集合 $\{M_k\}$ 。以 S_k 表示具有掩蔽器 M_k 的增强声景。换言之, S_k 是添加 M_k 后的声景 S_0 ,其中考虑播放系统和现实世界环境的响应。找到最优的 $\tilde{M}$ 以将对增强声景的给定感知(或情感属性) f 的评价最大化,相当于找到 $\operatorname{argmax}_k f(S_k)$,其中 f 可以是例如愉悦度、事件性或其他感知属性在数字指标上的评分。然后,我们可以用 $\tilde{M}$ 增强 S_0 ,以获得 $\tilde{S}$ 。

[0119] 例如,AMSS的朴素实施可以使用 $f(\cdot)$ 的某个或某种近似器,以为每个增强声景 S_k 输出感知属性的预测值 $\hat{f}_k := \hat{f}(S_k)$,并且选取具有最高 $\hat{f}_k$ 的掩蔽器。然而,如前所述,该确定性输出方案忽略与感知属性指标上的主观评分相关的、固有的不确定性。例如,一个或一种声景对一些人而言可能是非常愉悦的,但对另一些人而言却可能是非常烦扰的,与此同时,另一声景对任何听众而言可能几乎普遍地是略微愉悦的,这取决于情境。

[0120] 为收集该不确定性, f 可以被视为具有与某个或某种“随机”声景 S 的联合分布 $p(f, S)$ 的随机变量,其中观察到的增强声景 S_k 被视为 S 的现实化(realization)。然后,不确定性将以 f 的方差(variance)形式表示。在该情况下,预测模型将尝试将条件分布 $p(f|S = S_k)$ 建模为增强声景 S_k 的某个或某种函数 $\hat{F}(S_k)$ 。为简洁起见, $\hat{F}(S_k)$ 被缩写为 $\hat{F}_k$ 。

[0121] 神经近似器806可以用于输出 $\hat{F}_k$ 。神经近似器806可以称为概率感知属性预测器(probabilistic perceptual attribute predictor,PPAP)。在图8所示的、所提出的AMSS中,每个掩蔽器均可传递至前述PPAP 806,PPAP 806输出某个或某种预定义分布族的一个或多个参数。PPAP 806可以被训练以输出正态分布 $N(\mu_k, \sigma_k^2)$ 的平均值 μ_k 和对数标准差 $\log \sigma_k$,平均值 μ_k 和对数标准差 $\log \sigma_k$ 被用于对输出 $\hat{F}_k$ 进行建模。一旦计算出足够的(或全部的)预测分布 $\hat{F}_k$,便可基于某个或某种预定标准选择“最佳”掩蔽器。例如,可以仅基于最高的 μ_k 选择掩蔽器,可从 $\{\hat{F}_k\}$ 中采样相似于或类似于 $\{\hat{f}_k\}$ 的确定值的集合以激励掩蔽器探索,或者可以使用考虑 σ_k 的、较复杂的标准。

[0122] 尽管只有 f 的观测值是可用的,并且 f 的真实分布是不明的,但是仍然可以在给定输出分布的情况下,通过将基准真实值(ground truth)的对数概率以受贝叶斯优化

(Bayesian optimization) 启发的方式最大化,从而优化模型。在给定声景 S_0 、掩蔽器的集合 $\{M_k\}$ 和基准真实值 $\{f(S_k)\}$ 的情况下,对声景的损失函数的贡献可以由下式给出:

$$[0123] \quad \mathcal{J}_{prob.}(\{M_k\}; \{f(S_k)\}) = -\frac{1}{K} \sum_k \mathcal{L}(f(S_k); \mu_k, \sigma_k) \quad (1)$$

$$[0124] \quad = \frac{1}{K} \sum_k \left[\frac{1}{2} \left(\frac{f(S_k) - \mu_k}{\sigma_k} \right)^2 + \log \sigma_k \right], \quad (2)$$

[0125] 其中 $\mathcal{L}$ 是输出分布 $N(\mu_k, \sigma_k^2)$ 的对数密度函数,其中省略加性常数。在训练过程中,可以通过具有可用基准真实值的、批量的声景-掩蔽器对,以优化模型。

[0126] 如计算式(2)所示,使用正态分布的损失函数可以被认为是由对数标准差正则化(regularized)的加权MSE的损失。该损失函数相对于 σ_k 是固有地稳定的,因为第一项激励(encourage)较大的 σ_k ,与此同时,第二项激励较小的 σ_k 。当然,输出分布的一些其他选择可以类似地缩减(reduce)为其他偏差指标(deviation measures),例如缩减为正则化的平均绝对误差的拉普拉斯分布(Laplace distribution)。应注意的是,也可以通过将 σ_k 设置为某个预定常数且使用基准(ground)与 μ_k 之间的、纯MSE的损失训练模型,以强制模型进行确定性学习。为了进行消融研究(ablation study)的目的,可以使用下式

$$[0127] \quad \mathcal{J}_{det.}(\{M_k\}; \{f(S_k)\}) = \frac{1}{2K} \sum_k (f(S_k) - \mu_k)^2 \quad (3)$$

[0128] 作为计算式(2)的确定性对应项。计算式(3)中的确定性损失可以被认为是具有静态 $\sigma_k=1$ 的计算式(2)。

[0129] 验证实验

[0130] 为验证所提出的系统,我们令 f 为ISO12913-3标准定义的归一化的愉悦度量度, f 也可以称为ISO愉悦度。具体而言,

$$[0131] \quad f(S_k) = \frac{2(r_{pl}-r_{an})+\sqrt{2}(r_{ca}-r_{ch}+r_{vi}-r_{mo})}{8(1+\sqrt{2})} \in [-1,1] \quad (4)$$

[0132] 其中 $r_{pl}, r_{an}, r_{ca}, r_{ch}, r_{vi}, r_{mo} \in \{1,2,3,4,5\}$ 分别是以5点李克特量表表示的、参与者认为增强声景 S_k 愉悦、烦扰、平静、混乱、活跃和单调的程度。对于每个 S_k ,可以预测分布 $\hat{F}_k = N(\mu_k, \sigma_k^2)$,其中“基准真实值”标签 $f(S_k)$ 是参与者给出的、对 S_k 的ISO愉悦度评分。作为消融研究,将使用所提出的方法所训练的模型与通过 μ_k 直接预测 $f(S_k)$ 的确定性模型进行比较。验证实验是在对各个增强声景的主观反应的数据集上进行的。

[0133] 数据集

[0134] 该数据集包含5折交叉验证集(5-fold cross-validation set)中的、对增强声景的主观反应的、12600个去重集合(unique sets)以及独立测试集(independent test set)中的反应的48个附加集合,总共12648个样本。在给定对ISO12913-2情感反应问卷的以下6项子集的反应的集合:“您在多大程度上同意或反对当前周围声音环境是{愉悦的、混乱的、活跃的、平静的、烦扰的、单调的}?”的情况下,每个样本将增强声景(作为原始音频记录)映射于ISO愉悦度值。

[0135] 参与者按照5分量表进行回答,其中该5分量表具有多个标签(labels)“非常反对”、“反对”、“既不同意也不反对”、“同意”和“非常同意”,该多个标签分别编码为1、2、3、4和5。然后,将编码后的回答用于计算式(4)中,以为每个增强声景计算ISO愉悦度的基准真

实值标签。

[0136] 增强声景

[0137] 5折交叉验证集中的增强声景是通过将Freesound和xeno-canto的录音的30秒摘录作为“掩蔽器”添加于世界城市声景 (Urban Soundscapes of the World, USotW) 数据库的声景的立体声录音 (binaural recordings) 的30秒摘录以制成的, 世界城市声景数据库是全面的城市声景数据集。也包括没被增强的声景以作为对照。所使用的全部录音均以44.1kHz采样。

[0138] 每一折具有56个掩蔽器的组, 该56个掩蔽器的组分为以下类别: 鸟 (16)、施工 (8)、交通 (8)、水 (16) 和风 (8)。选择这些类别以涵盖总体上被评估为愉悦的和烦扰的声音类型范围。每一折中的掩蔽器和声景都是互斥的 (disjoint)。

[0139] 测试集中的增强声景以相似的方式制成, 其中7个掩蔽器 (包括没被增强的对照掩蔽器在内共8个掩蔽器) 独立于交叉验证集中的掩蔽器, 并且声景的6个立体声录音独立于USotW数据集, 该6个立体声录音使用相同的声景索引协议 (Soundscape Indices Protocol) 录制。7个掩蔽器摘录自xeno-canto音轨ID (识别号) 640568 (鸟“B1”) 和568124 (鸟“B2”), 以及Freesound音轨ID586168 (施工“Co”)、587219 (交通“Tr”)、587000 (水“W1”)、587759 (水“W2”) 和587205 (风“Wi”)。掩蔽器被完全地 (exhaustively) 添加于测试集的全部声景, 这在测试集中产生48个样本。在为测试集以恒定的、0dB的声景与掩蔽器的比率以及为交叉验证集以dB为单位、从 $\{-6, -3, 0, 3, 6\}$ 中随机选定的值添加掩蔽器之前, 全部声景均根据K. Ooi等人描述的方法 (“在人造头上进行立体声耳机音频校准的自动化”, MethodsX, 第8卷, 第2期, 第101288页, 2021年) 校准为实时测量的A加权等效的SPL ($L_{A,eq}$)。

[0140] 主观反应

[0141] 为获得对增强声景的主观反应, 招募300名参与者, 每人对从验证集的一个折中随机选择的42个独特的增强声景进行评分。招募另外5名参与者, 每人对测试集中的48个增强声景进行评分。因此, 验证集和测试集中的每个增强声景分别由一名参与者和五名参与者评分。所有参与者都使用一副由外部声卡 (Creative Sound Blaster E5) 驱动的耳罩式耳机 (Beyerdynamic Custom One Pro) 以收听经校准的增强声景。听完每个增强声景后, 他们回答前文描述的6项问卷 (在“数据集”标题下), 并且根据计算式4从他们的回答中计算出ISO愉悦度。

[0142] 模型与训练

[0143] 增强声景的对数梅尔频谱图 (Log-mel spectrograms) 被用作输入。使用具有50%重叠率的4096样本汉恩窗口 (Hann window) 提取对数梅尔频谱图, 并且将该对数梅尔频谱图压缩至64个梅尔箱 (mel bins)。图9示出 (a) 根据各个实施例使用的基本卷积循环神经网络 (CRNN) 架构; (b) 根据各个实施例, (a) 中的特征映射块的一个可能实施; 以及 (c) 根据各个其他实施例, (a) 中的特征映射块的其他可能实施。研究四个不同的特征映射块, 即使用双向门控循环单元 (bidirectional gated recurrent unit, BiGRU) 的基础映射块 (vanilla mapping block)、加性注意力块 (additive attention block)、点积注意力块 (dot-product attention block) 和具有4个头的多头注意力块。基础映射块如图9 (b) 所示, 与此同时, 共享相同通用工作流程的、全部的基于注意力的块如图9 (c) 所示。图9 (a) 的模型的最后数层是全连接层 (dense layers), 该全连接层最终输出 μ_k 和 $\log \sigma_k$ 。在确定性消

融模型中, $\log \sigma_k$ 被忽略。

[0144] 全部模型均使用5折交叉验证方案进行训练,其中每一折具有10个模型,总共50个模型。每一折为多个模型使用相同的10个种子。全部模型均使用亚当优化器(Adam optimizer)以 5×10^{-5} 的学习率训练最多100轮。对于每个模型,在交叉验证集和测试集二者中均使用具有最佳验证损失的模型权重进行评估。

[0145] 结果与讨论

[0146] 图10是示出根据各个实施例的概率感知属性预测器(PPAP)在针对每个设置测试的10次运行中的均折均方误差(mean fold mean squared errors, MSEs) ($\pm$ 标准差) (上:交叉验证集,下:测试集)。每次运行可以包括五个模型,每个模型在交叉验证集的不同折上,但是具有相同的初始条件。对于全部模型,对MSE的贡献被计算为 $(\mu_k - f_k)^2$ 。星号(*)表示统计上显著的改进 ($p < 0.05$)。图10总结前文描述的验证实验在5折交叉验证集和独立测试集二者上的结果。作为参考,还提供一个简单的(trivial)确定性“标签均值模型”的结果。在该“标签均值模型”中,训练集中的标签均值用作验证集和测试集中全部刺激(stimuli)的预测。所有其他研究模型的表现均优于标签均值模型,因此表明由其他研究模型进行的特征提取是有益的。这可能是因为,与验证集(12600个样本、280个掩蔽器)相比,测试集较小且多样性较低(48个样本、7个掩蔽器)。

[0147] 此外,对于所测试的四个架构,具有概率输出的模型表现优于具有验证集和测试集二者的确定性输出的等同模型。图10还示出每个模型的MSE降低百分比。可以观察到,具有多头注意力块的CRNN和基础CRNN在验证集和测试集的MSE中分别呈现1.1%和7.8%的最大改进。

[0148] 为使这些降低的显著性(significance)量化,对验证集MSE和测试集MSE二者进行确定性模型与概率模型之间的双侧威尔科克森符号秩测试(two-sided Wilcoxon signed-rank tests)。在图10中,在5%显著性级别上显著的降低以星号标记。对于基础CRNN和具有加性注意力的PPAP,测试集MSE的降低在统计上是显著的,并且提供证据以支持所提出的方法对没有看到的数据的泛化性(generalizability)。这可以归因于测试集参与者贡献的固有随机性与训练集中的参与者不同,并且PPAP对该随机性的泛化较好。然而,需要进一步的消融研究以验证该假设。

[0149] 然而,测试集MSE的最低改进实际上是在具有多头注意力块的PPAP中观察到的。这可以归因于具有多头注意力块的PPAP比其他三个模型具有多出约50%的参数(约120K对约80K)的事实,这可能导致训练模型(确定性训练模型和概率性训练模型二者)过度拟合于相对较小的12600个训练和验证样本的数据集。

[0150] 图11示出使用测试集上的50个模型(点积变量)进行掩蔽选择,其中采用根据各个实施例的朴素最大 μ_k 选择方案(左)以及根据各个实施例的随机采样方案以激励掩蔽器探索(右)。在每个子图的顶部上示出用于每个基本声景的选择标准。单元格颜色表示测试集参与者评定的平均ISO愉悦度。圆圈尺寸表示选择该掩蔽器的模型数量。掩蔽器描述已在前文提供(在标题“增强声景”下)。

[0151] 在朴素选择方案中,大多数模型在全部基本声景中选择第一鸟鸣掩蔽器(B1)。这与先前声景研究中的共识一致,该先前声景研究总体上观察到对鸟鸣掩蔽器的、较高愉悦度的反应。然而,第二鸟鸣掩蔽器(B2)可能总体上获得参与者的较低评分,并且相应地,模

型的投票数较低,这是因为它在测试集声景的情境中被认为愉悦度较低,这在先前的研究中已经观察到。

[0152] 在随机选择方案中,系统仍然倾向于选择为大多数声景提供良好的愉悦度得分的B1,但是现在,系统对其他掩蔽器的探索比朴素方案多得多。这在实时系统中可以是有益的,在实时系统中可以实时地获得人员反馈,以自适应地调整掩蔽选择或改进未来的模型。还可以看出,在声景5的情况下,没被增强的声景的愉悦度已经相对较高,探索率高于其他基本声景,这允许在愉悦度级别上仅做出较小妥协的情况下,提供较多样化的声学体验。

[0153] 结论

[0154] 各个实施例可以涉及用于以人为本的城市声景增强的自动掩蔽选择系统(AMSS),该系统使用概率感知属性预测器(PPAP)。所提出的PPAP是使用卷积循环神经网络实施的,并且经过训练从而以概率方式输出预测。这允许该系统在考虑声学刺激的人类主观感知中固有的随机性的同时,预测声景的感知属性。通过一项有超过300名参与者和超过12000个独特声景的大规模收听测试,我们验证了我们的PPAP在预测增强声景的愉悦度方面的有效性,该增强声景包括从没被看到的声景和掩蔽器中生成的声景。AMSS的未来工作可能包括实时实施以评估其生态效度(ecological validity),以及对所提出的方法对例如事件性或平静度的其他感知属性(f)的研究,这是因为所提出的方法不是专用于任何特定属性的。事实上,由于PPAP所依据的主要假设是随机的基准真实值标签的主要假设,人们也可以设想将其应用于任何需要预测主观评价的情况。

[0155] 涉及对给定声景添加称为“掩蔽器”的声音的声景增强是以人为本的城市噪声缓解措施,该措施旨在提高整体声景质量。然而,掩蔽器的选择通常是基于繁琐的过程预测的,并且对于现实世界声景的时变性而言是不灵活的。由于每个声景的感知独特性以及人类感知固有的主观性,本申请提出了概率感知属性预测器(PPAP),该概率感知属性预测器预测随机分布的参数作为输出,以代替单个确定值作为输出。在使用PPAP的情况下,本申请开发了自动掩蔽选择系统(AMSS),该系统基于用于给定声景的ISO12913-3愉悦度评分的预测分布,选择最佳掩蔽器候选。通过有300名参与者的的大规模收听测试,收集了12600个主观反应,每个主观反应针对一个独特的增强声景,从而以5折交叉验证方案训练PPAP模型。在使用卷积循环神经网络主干以及具有为PPAP进行的注意力机制的数个变量的测试的情况下,本申请使用具有48个没被看到的增强声景的盲测集,进行了对所提出的系统的评估,以评价概率输出方案相对于常规确定性系统的有效性。

[0156] 示例研究2

[0157] 介绍

[0158] 城市噪声污染的减弱是复杂且多方面的问题,其中噪声源的消除或减弱通常不是切实可行的选择。与旨在降低声压级或声能的常规噪声控制方法不同,声景方法提出基于感知声学舒适度的、较全面的、以人为本的策略。这种通常称为声景增强的技术涉及将“所期望出现的”声音添加于声学环境中,以“掩蔽”噪声且改善整体感知声学质量,这种技术在虚拟环境和现实环境二者中,针对各个类型的噪声源都取得了令人满意的结果。

[0159] 然而,“掩蔽器”和播放级别的选择常规上是任意的、专家指导的,或者是基于事后分析的。这些掩蔽选择过程不仅通常是耗时耗力的,而且对于现实世界声景的动态特性而言是不灵活的。因此,这些掩蔽器和播放级别的静态选择通常不可能始终实现最佳的声学

舒适度。为解决这些限制,示例研究1示出通过由概率深度学习模型驱动的掩蔽选择系统的使用,实现的适应性声景增强的一个首次尝试,该概率深度学习模型为特定增强声景预测“愉悦度”的概率分布。

[0160] 尽管上述概率感知属性预测器 (PPAP) 模型为深度学习驱动的自动声景增强提供令人信服的概念证明,但是该模型可能不是为实时部署设计的。该PPAP模型的实施需要将增强声景的预混音轨作为输入。混合过程是在相对于所录制的没被增强的声景音轨的已知的实时SPL级别、以预先校准的声景与掩蔽器的比率完成的。这可能会导致无视掩蔽增益级别的模型--掩蔽增益级别是实时部署的关键播放参数。此外,要求将基本声景记录与掩蔽器的混合作为输入,给整个系统增加不必要的复杂性。首先,全部候选掩蔽器都需要播放所需的、处于特定声景与掩蔽器的比率 (SMR) 的数字增益查对表。其次,在选择过程中,每个候选掩蔽器-SMR对都需要计算增强声景音轨,这增加了系统的计算成本 (computational overhead)。

[0161] 各个实施例可以呈现具有修改的计算式的PPAP模型,该PPAP模型将允许在现实世界的部署中实现较高效的计算和带宽系统。所提出的模型可以在输入阶段将基本声景、掩蔽器和增益级别相互解耦,并且替代地为基本声景和掩蔽器引入单独的特征提取器分支。通过使用掩蔽器的、独立于实时声压级 (SPL) 的数字增益级别作为调节性输入,“增强”过程仅发生于特征空间以代替波形空间。然后基于基本声景特征、掩蔽特征和增益调节特征,使用注意力机制以“查询”目标感知属性的分布。所提出的方法导致系统多个组件的复杂性降低。首先,可以不需要将声景的原始波形发送至推理引擎以与掩蔽器混合,因为可以使用带宽消耗较少的频谱数据作为模型输入。对于云推理,这意味着边缘的数据流出率 (data egress rate) 的显著降低。其次,掩蔽特征可以独立于增益级别进行预计算,这减少了推理时间,并且因此减少了总体延迟 (latency)。第三,现在可以在无需重新计算增强声景或甚至声景和掩蔽特征的情况下,较快地对同一掩蔽器的多个增益级别进行属性预测。最后,现在可以在无需创建查对表的情况下,实现新掩蔽器对系统的添加。

[0162] 提出的方法

[0163] 准备工作

[0164] 考虑数字化声景记录 $\mathbf{s}_i[n] \in [-1,1]^C$, 其中 $C \geq 1$ 是通道数, 并且 n 是样本索引。通过 $a_i = \mathcal{A}(\mathbf{s}_i)$ 表示 $\mathbf{s}_i$ 的、以 dBA 为单位的实时 A 加权等效声压级 $L_{A,eq}$ 。为简洁起见, 所有提到的声压级 (SPL) 均指 30 秒 $L_{A,eq}$, 除非另有说明。假设所有声景均由相同的实时记录设置记录, 并且在假设近似线性的情况下, 声压与数字级别的比率 (sound pressure-to-digital level ratio, SPDR), 即线性尺度上的相对声压级与设置的数字满量程 (digital full scale, DFS) 的比率, 可以认为大致恒定, 并且以 d_0 表示, 其单位是 $Pa/(p_0 \cdot DFS)$, 其中 $p_0 = 20 \mu Pa$ 是基准声压级。由于全部声景输入都源自同一录音设备, 这也意味着实时 SPL 信息隐含地嵌入输入频谱图数据中。

[0165] 考虑单通道掩蔽器 $\mathbf{m}_j[n] \in [-1,1]$ 。对于掩蔽器, 假设它们不都是由同一录音装置录制的。在实践中, 这是因为掩蔽器通常源自与实时录音装置不同的录音系统, 并且通常通过例如 Freesound 和 Xeno-canto 的开放内容供应商获得。因此, 每个掩蔽器的 SPDR 通常是不明的 (unknown), 并且不能假设其与任何其他掩蔽器的 SPDR 相同。以 d_j 表示掩蔽器 $\mathbf{m}_j$ 的

SPDR。在(通过对本文中使用的全部掩蔽器进行彻底的增益校准以验证的)近似线性的假设下,可以通过将掩蔽器 m_j 以 d_0/d_j 定比例,以将掩蔽器 m_j 归一化为与声景录音相同的SPDR。如果 d_j 和实时SPL a_i 二者是已知的,则在无需进一步校准的情况下,具有特定声景与掩蔽器的比率(SMR)的、以dBA为单位的声景增强可以相对容易地进行。然而, d_j 在实践中通常是不明的,因此常规解决方法是在使用具有 d_0^{-1} 的数字级别与声压的比率(digital level-to-sound pressure ratio, DSPR)的播放系统的情况下,进行校准,以在 l 的多个值中,获得以 l 的声压级再现 m_j 所需的数字增益 $g_j(l)$ 。获得 $g_j(\cdot)$ 的精确值所需的校准通常可能需要专门的设备,例如具有人造头的录音装置,并且可能是耗时的和/或耗力的,这取决于所使用的校准方法。在本文中, g_j 明确用作调节性输入,因此去除了SPL相关的SMR的需求或掩蔽校准的需求。

[0166] 模型

[0167] 图12A示出根据各个实施例的概率感知属性预测器(PPAP)的示意图。设 $S_i \in \mathbb{R}^{T \times F \times C}$ 为声景 s_i 的对数梅尔频谱图,并且 $M_j \in \mathbb{R}^{T \times F \times 1}$ 是掩蔽器 m_j 的对数梅尔频谱图,其中 T 是频谱图时间帧的数量,并且 F 是梅尔箱的数量。在本文中, $C = 2, F = 64$,并且 $T = 644$ 对应30秒的信号,该信号以44.1kHz、以4096个样本的短时傅里叶变换(short-time Fourier transform, STFT)窗口尺寸以及50%的帧重叠率采样。声景和掩蔽嵌入分别由特征提取器1222 f_s 和特征提取器1226 f_m 计算,以使得

$$[0168] \quad k_i = f_s(S_i) \in \mathbb{R}^{N \times D}, \quad (5)$$

$$[0169] \quad q_j = f_m(M_j) \in \mathbb{R}^{N \times D} \quad (6)$$

[0170] 其中 $D = 128$ 是嵌入维度,并且 $N = 20$ 是特征时间帧的数量。每个特征提取器1222、1226可以包括5个卷积块,每个卷积块包含具有 3×3 的核和1步幅的卷积层;批量归一化;具有概率0.2的退出率(dropout);Swish激活;以及 2×2 的平均池化层。卷积层按顺序包含16、32、48、64和64个输出通道。

[0171] 在使用注意力的查询-键-值(query-key-value, QKV)模型的情况下,感知属性预测过程可以看作是使用掩蔽器查询映射,其中没被增强的声景给出键,并且增强声景给出值。我们考虑在嵌入域中执行“增强”(在特征增强块1230中),以代替在音频或频谱域中执行声景增强。这可以通过使用以掩蔽数字增益调节的简单映射完成,以使得

$$[0172] \quad v_{i,j,g} = f_g(k_i, q_j, \gamma) \in \mathbb{R}^{N \times D}, \quad (7)$$

[0173] 其中 g 是查询增益级别, $\gamma = \log_{10} g$, f_g 是增益调节增强层。有 f_g 的三个不同的实施,即

$$[0174] \quad f_g^{(CAT)}(k_i, q_j, \gamma) = \text{Dense}^2(k_i | q_j | \gamma I^{N \times 1}) \quad (8)$$

$$[0175] \quad f_g^{(ADD)}(k_i, q_j, \gamma) = \text{Dense}^2(k_i + \gamma q_j) \quad (9)$$

$$[0176] \quad f_g^{(CONV)}(k_i, q_j, \gamma) = \text{Dense}(\text{Conv}(k_i \| q_j \| \gamma I^{N \times D})) \quad (10)$$

[0177] 其中张量 $|$ 是最后一个轴上的张量拼接, $\|$ 是创建附加轴的张量堆叠运算符, $\text{Dense}: \mathbb{R}^{N \times \bullet} \mapsto \mathbb{R}^{N \times D}$ 是具有 D 个输出单元的全连接层, $\text{Conv}: \mathbb{R}^{N \times D \times 3} \mapsto \mathbb{R}^{N \times D}$ 是具有

D个过滤器的、使用一维核将堆叠的维度压缩为单一轴的卷积层。图12B示出根据各个实施例中的计算式(8)、(9)和(10)的三种特征增强方法。在训练过程中,如果掩蔽器 m_j 是静音音轨(即无增强),我们将使 $\gamma \sim N(\nu, \zeta^2)$ 随机化,其中 ν 和 ζ 是具有非静音掩蔽器的训练样本的对数增益的平均值(mean)和标准差(standard deviation)。这是为了教导模型忽略静音掩蔽器的增益值。在查询、键和值嵌入到位的情况下,可以使用任何QKV兼容的注意力块1220a f_a ,计算输出嵌入,以使得

$$[0178] \quad \mathbf{z}_{i,j,g} = f_a(\mathbf{q}_j, \mathbf{k}_i, \mathbf{v}_{i,j,g}) \in \mathbb{R}^D. \quad (11)$$

[0179] 然后,输出块1220b f_o 计算预测分布 $\hat{\mathbf{Y}}_{i,j,g}$,输出块1220b f_o 预测分布 $\hat{\mathbf{Y}}_{i,j,g}$ 的参数 $\mu_{i,j,g}$ 和 $\log \sigma_{i,j,g}$ 。可以考虑三个类型的注意力,即加性注意力(additive attention, AA)、点积注意力(dot-product attention, DPA)和多头注意力(multi-head attention, MHA)。在该示例中,使用4头MHA。

[0180] 损失函数

[0181] 如同示例研究1,人类主观反应可以被认为是固有地随机的,一方面,参与者是总体的一个样本,另一方面,每个参与者在收听测试过程中给出的反应也是他们固有的非确定性感知的随机样本。因此,用于目标属性的标签 $\mathbf{y}_{i,j,g}$ (即ISO愉悦度)可以被视为代表目标属性分布的随机变量 $\mathbf{Y}_{i,j,g}$ 的观察结果,目标属性的真实分布是不明的。在使用最大似然计算式的情况下,PPAP模型的优化目标可以通过以下计算式给出:

$$[0182] \quad \argmax_{\theta} \sum_{i,j,g} \mathcal{L}_{\hat{\mathbf{Y}}_{i,j,g}}(\mathbf{y}_{i,j,g}) \quad (12)$$

[0183] 其中 θ 是模型的参数,并且 $\mathcal{L}_x(x)$ 是在x处估算的X的对数密度。通过独立的正态分布对每个声学场景的主观响应进行建模,计算式(12)中的优化目标转换为损失函数

$$[0184] \quad \mathcal{T} = \frac{1}{|\mathfrak{T}|} \sum_{i,j,g} \frac{1}{2} \left(\frac{\mathbf{y}_{i,j,g} - \mu_{i,j,g}}{\sigma_{i,j,g}} \right)^2 + \log \sigma_{i,j,g}, \quad (13)$$

[0185] 其中 $\mathfrak{T}$ 是训练轮次。

[0186] 优化推理

[0187] 尽管模型在训练过程中看到不同声景-掩蔽器-增益样本的多个批次,但是在实时推理期间,模型每次只看到一个基本声景,与此同时,经历多个掩蔽器和增益级别以为特定声景找到最合适的掩蔽器-增益对。因此,通过消除重复计算,以代替为每个掩蔽器-增益组合进行新的前向传递,可以获得显著的性能改善。分别以 τ_s 、 τ_m 和 τ_{g+a+o} 表示执行 f_s 、 f_m 和 $f_g \circ f_a \circ f_o$ 所需的时间。设 η_m 为每个声景待查询的掩蔽器的数量,并且 η_g 为每个掩蔽器待查询的增益值的数量。朴素批处理方案将需要 $\eta_g \eta_m (\tau_s + \tau_m + \tau_{g+a+o})$ 的总运行时间。相比之下,通过优化查询过程,如图13的算法1所示,总运行时间可以减少至 $\tau_s + \eta_m \tau_m + \eta_g \eta_m \tau_{g+a+o}$ 。图13示出根据各个实施例的、减少总运行时间的算法。附加地,对于固定的掩蔽器组,可以预先计算掩蔽特征 $\mathbf{q}_j$,这允许将总运行时间进一步减少至 $\tau_s + \eta_g \eta_m \tau_{g+a+o}$ 。

[0188] 数据集

[0189] 本文中使用的数据集从示例研究1中使用的数据集扩展有额外的参与者,共计442名参与者的18564个数据点(每位参与者42个刺激)。每个数据点代表参与者对增强声景的

主观反应。

[0190] A. 刺激

[0191] 通过将30秒的“掩蔽器”添加至源自世界城市声景 (USotW) 数据库的声景的30秒立体声景录音, 增强声景在5个互斥的折中制成。也包括没被增强的声景 (即使用静音掩蔽器“增强”的声景) 以作为对照。所使用的全部录音以44.1kHz采样。更多详细信息可以在示例研究1中找到。

[0192] 将掩蔽器添加至声景是在 $\{-6, -3, 0, 3, 6\}$ dBA的、特定的声景与掩蔽器的比率 (SMR) 下制成的, 其中SMR由声景的SPL与掩蔽器的SPL的比率定义。虽然已知基本声景音轨的实时SPL, 但是掩蔽器数据源自在线存储库, 特别是Freesound和Xeno-canto, 因此掩蔽器数据需要校准。通过在虚拟头 (dummy head) (GRAS 45BB KEMAR Head & Torso) 上以1dBA步长, 对音轨进行彻底校准至46dBA至83dBA的闭区间中的SPL值, 以创建将第 j 个掩蔽器放大至特定SPL所需的数字增益 $g_j(l)$ 的查对表。校准过程是半自动的, 其中使用与参与者相同的音频播放设备。由于基本声景的实时SPL通常是非整数的, 非整数 l 的数字增益是从查对表插值得到的, 其中使用

$$[0193] \quad g_j(l) = \text{lookup}_j[\text{round}(l)]. 10^{(l - \text{round}(l))/20} \quad (14)$$

[0194] 其中 $\text{lookup}_j[\lambda]$ 是在 λ dBA的SPL下用于掩蔽器 j 的校准增益。

[0195] B. 主观反应

[0196] 全部参与者都使用一副环绕式耳机 (Beyerdynamic Custom One Pro) 收听经校准的增强声景, 该耳机由外部声卡 (Creative SoundBlaster E5) 驱动。收听完每个增强声景后, 参与者回答一份基于ISO12913-2标准的6项问卷: “您在多大程度上同意或反对当前周围声音环境是[...]?”, 其中[...]是愉悦的、混乱的、活跃的、平静的、烦扰的和单调的中的一个。每个问卷项目都有“非常反对”、“反对”、“既不同意也不反对”、“同意”和“非常同意”的多个选项, 该多个选项被数字编码为从1至5。然后通过以下计算式, 计算每个增强声景的ISO愉悦度的基准真实值标签

$$[0197] \quad y_{i,j,g} = \frac{2(v_{pl} - v_{an}) + \sqrt{2}(v_{ca} - v_{ch} + v_{vi} - v_{mo})}{8(1 + \sqrt{2})} \in [-1, 1] \quad (15)$$

[0198] 其中 v_{pl} 、 v_{an} 、 v_{ca} 、 v_{ch} 、 v_{vi} 、 v_{mo} 分别是愉悦的、烦扰的、平静的、混乱的、活跃的和单调的的问卷项目的数字评分。

[0199] 测试

[0200] 对于本文中的全部模型类型, 每个模型都以5折交叉验证的方式, 为每个验证集的相同的10个种子, 进行训练, 每个模型类型总共50个模型。使用具有 5×10^{-5} 的学习率的亚当优化器, 对每个模型进行最多100轮的训练。如果验证均方误差 (MSE) 在至少5轮内没有下降, 则学习率减半, 并且如果验证MSE在至少10轮内没有下降, 则提前停止训练。本文披露的MSE以及平均绝对误差 (MAE) 是使用预测分布的平均值 $\mu_{i,j,g}$ 与基准真实值标签 $y_{i,j,g}$ 之间的差值计算的。

[0201] 除上文提到的三个注意力类型 (在标题“模型”下) 之外, 测试中还包括作为基线模型的“直通”注意力块。基线模型在图14中以“X”表示。直通块可以定义为

$$[0202] \quad \text{PassThroughAttention}(q_j, k_i, v_{i,j,g}) = v_{i,j,g} \quad (16)$$

[0203] 图14示出(上)均方误差(MSE)作为注意力块类型的函数的图示,该图示出根据各个实施例的每个注意力块类型和特征增强方法的愉悦度预测的验证MSE的小提琴图;以及(下)平均绝对误差(MAE)作为注意力块类型的函数的图示,该图示出根据各个实施例的每个注意力块类型和特征增强方法的愉悦度预测的验证MAE的小提琴图。图14示出根据每个注意力和增强类型的MSE和MAE的预测误差分布。在全部注意力类型中,CONV增强方法在MSE和MAE两个方面表现最佳,其次是CAT和ADD。这可以归因于CONV方法和内核过滤,CONV方法天然地激励 k_i 与 q_j 之间的特征对齐(feature alignment),与其他两个方法相比,内核过滤允许在特征空间中进行较灵活的增强操作。

[0204] 还可以看出,与无注意力的模型相比,使用QKV注意力总体上允许获得较好的性能。然而,只要使用注意力的某个或某种形式将查询、键和值张量耦接,注意力机制的类型似乎不对模型性能产生显著影响。尽管AA和DPA只有一个注意力头,但是四头MHA似乎没有在预测误差方面获得可察觉的改善。还应当注意,特征增强块的选择似乎对模型的准确性有较显著的影响,如所看到的无注意力的CONV增强模型总体上比有注意力的ADD模型表现好的事实。

[0205] 尽管模型从没看到呈现给参与者的、增强声景的音频数据,在 0.122 ± 0.005 处的最佳平均MSE模型(具有CONV增强的卢昂注意力(Luong attention))与示例研究1中的结果相当,示例研究1使用增强声景的对数梅尔频谱图作为输入。这可以证明特征增强技术在代替波形空间的特征空间中模拟声景增强的有效性。

[0206] 图15A是ISO愉悦度作为对数增益的函数的图示,该图示出根据各个实施例,使用具有CONV增强(种子5、验证折2)的加性注意力(AA),为掩蔽器类别“水”进行增益插值。图15B是ISO愉悦度作为对数增益的函数的图示,该图示出根据各个实施例,使用具有CONV增强(种子5、验证折2)的加性注意力(AA),为掩蔽器类别“交通”进行增益插值。图15C是ISO愉悦度作为对数增益的函数的图示,该图示出根据各个实施例,使用具有CONV增强(种子5、验证折2)的加性注意力(AA),为掩蔽器类别“鸟”进行增益插值。图15D是ISO愉悦度作为对数增益的函数的图示,该图示出根据各个实施例,使用具有CONV增强(种子5、验证折2)的加性注意力(AA),为掩蔽器类别“施工(construction)”进行增益插值。图15E是ISO愉悦度作为对数增益的函数的图示,该图示出根据各个实施例,使用具有CONV增强(种子5、验证折2)的加性注意力(AA),为掩蔽器类别“静音”进行增益插值。为实现该可视化的目的,全部掩蔽器均在零对数增益下调整为65.0dBA。基本声景具有65.32dBA的实时SPL级别。在训练过程中,模型看不到基本声景和全部掩蔽器。图15A至图15E中的实线表示预测分布的平均值。阴影区表示最高至预测平均值上方和最低至预测平均值下方的预测标准差的愉悦度级别。

[0207] 图15A至图15E可以通过为不同类型的掩蔽器在 $[-2, 2]$ 区间内、在256个对数增益值中查询ISO愉悦度分布,以展示模型的增益意识特性(gain-aware nature)。还包括静音“掩蔽器”音轨,以测试模型同时考虑掩蔽器和掩蔽器的增益级别二者的影响的能力。尽管愉悦度预测不是完全恒定的,可以看出,在不考虑查询增益值的情况下,模型都可以辨别出掩蔽器是静音的,并且输出大致恒定的愉悦度预测。对于其他非静音掩蔽器,可以看出,模型意识到每个掩蔽器的特性。与先前研究的发现一致,模型预测愉悦度随着鸟鸣掩蔽器增益的增加而增加,直到鸟鸣掩蔽器的增益接近环境级别($\gamma \approx -0.5$),然后愉悦度下降至没被增强的声景(即静音掩蔽器预测)的愉悦度以下。水声掩蔽器也呈现相似的效果,水声掩

蔽器的、最高至 $\gamma \approx -0.5$ 的愉悦度级别高于没被增强的声景的愉悦度,但是由于水声较连续的特性,愉悦度级别在比鸟鸣掩蔽器低的增益处开始下降。对于已知令人不愉快的掩蔽器,例如交通和施工,随着增益级别的提高,模型正确地输出越来越令人不愉快的预测。

[0208] 结论

[0209] 各个实施例可以涉及改进的概率感知属性预测器模型,该模型允许对增强声景的主观响应进行增益意识预测。所提出的模型将掩蔽器与声景特征提取解耦,并且在特征空间中模拟声景增强过程,从而消除在波形域中进行计算方面成本高昂的混合过程的需求。此外,该模型被重制以考虑掩蔽器的数字增益级别,以代替声景与掩蔽器的比率,这允许在无需耗时的校准的情况下,进行源自多个源的掩蔽器的使用。模型的模块化设计允许在推理时间内,进行大量特征的重用和预计算,这减少了部署所需的总体延迟和计算资源。在使用源自442名参与者的18000个主观响应的大规模数据集的情况下,已经证明了该模型能够准确预测与声景、掩蔽器和增益相关的愉悦度评分的能力。

[0210] 在声景增强系统中,掩蔽器和播放增益级别的选择对于声景增强系统在改善给定环境的整体声学舒适度的有效性而言是至关重要的。常规上,适当的掩蔽器和增益级别的选择是由可能不能代表目标人群的专家意见决定的,或者是由可能耗时耗力的收听测试决定的。此外,所获得的掩蔽器和增益的静态选择通常对于现实世界声景的动态特性而言是不灵活的。在本文中,深度学习模型被用于对给定声景的最佳掩蔽器及最佳掩蔽器的增益级别进行联合选择。采用高度模块化的构建块设计所提出的模型,这允许获得优化的推理过程,优化的推理过程可以快速搜索大量掩蔽器和增益的组合。此外,引入在数字增益级别方面调节的特征域声景增强的使用,这消除了推理期间计算方面的成本高昂的波形域混合过程,以及掩蔽器所需的、繁琐的预校准过程。所提出的系统在超过440名参与者对增强声景的主观响应的大规模数据集上得到验证,这确保了模型能够预测掩蔽器及掩蔽器的增益级别对感知愉悦度级别的综合影响的能力。

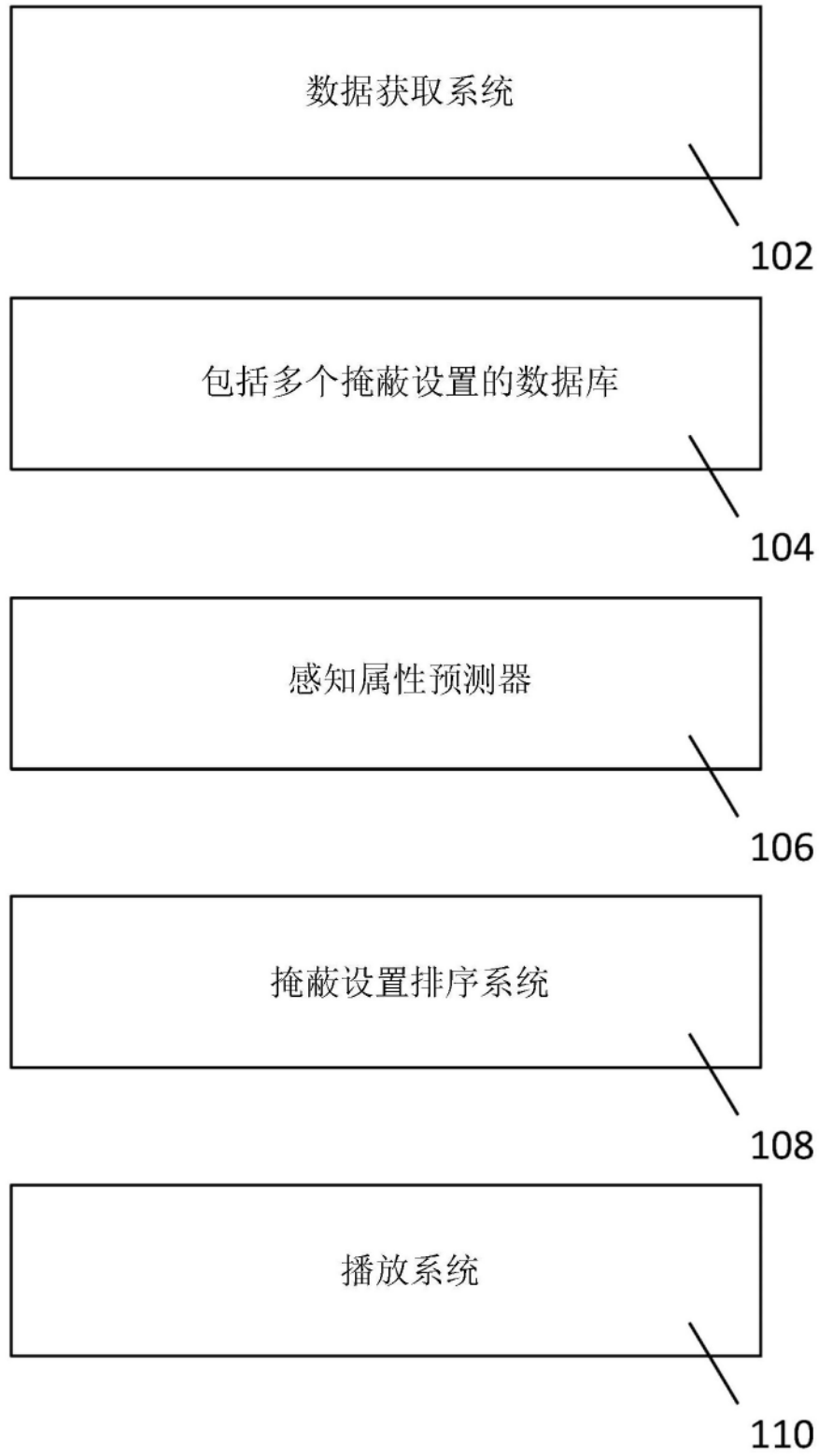


图1

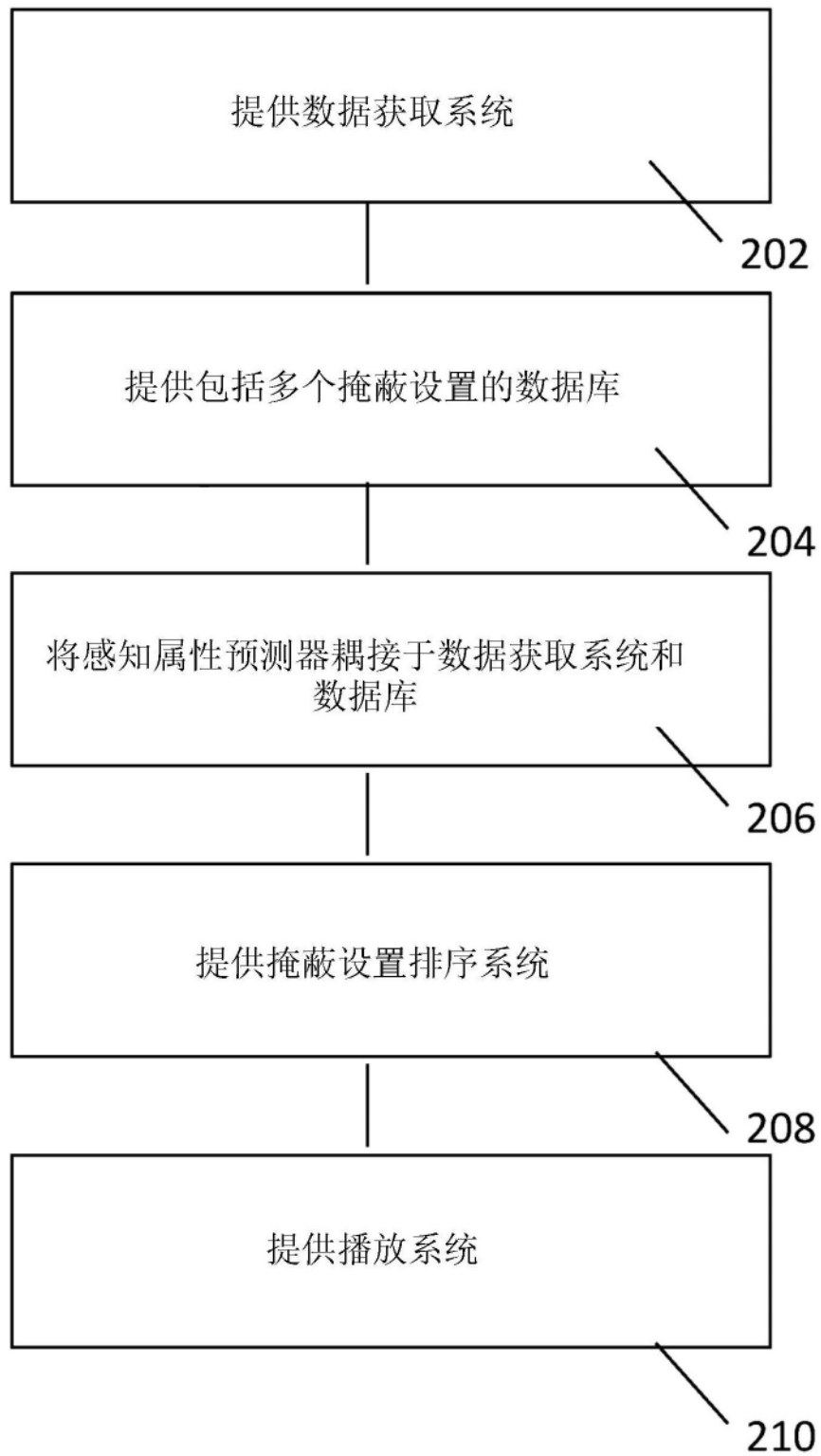


图2

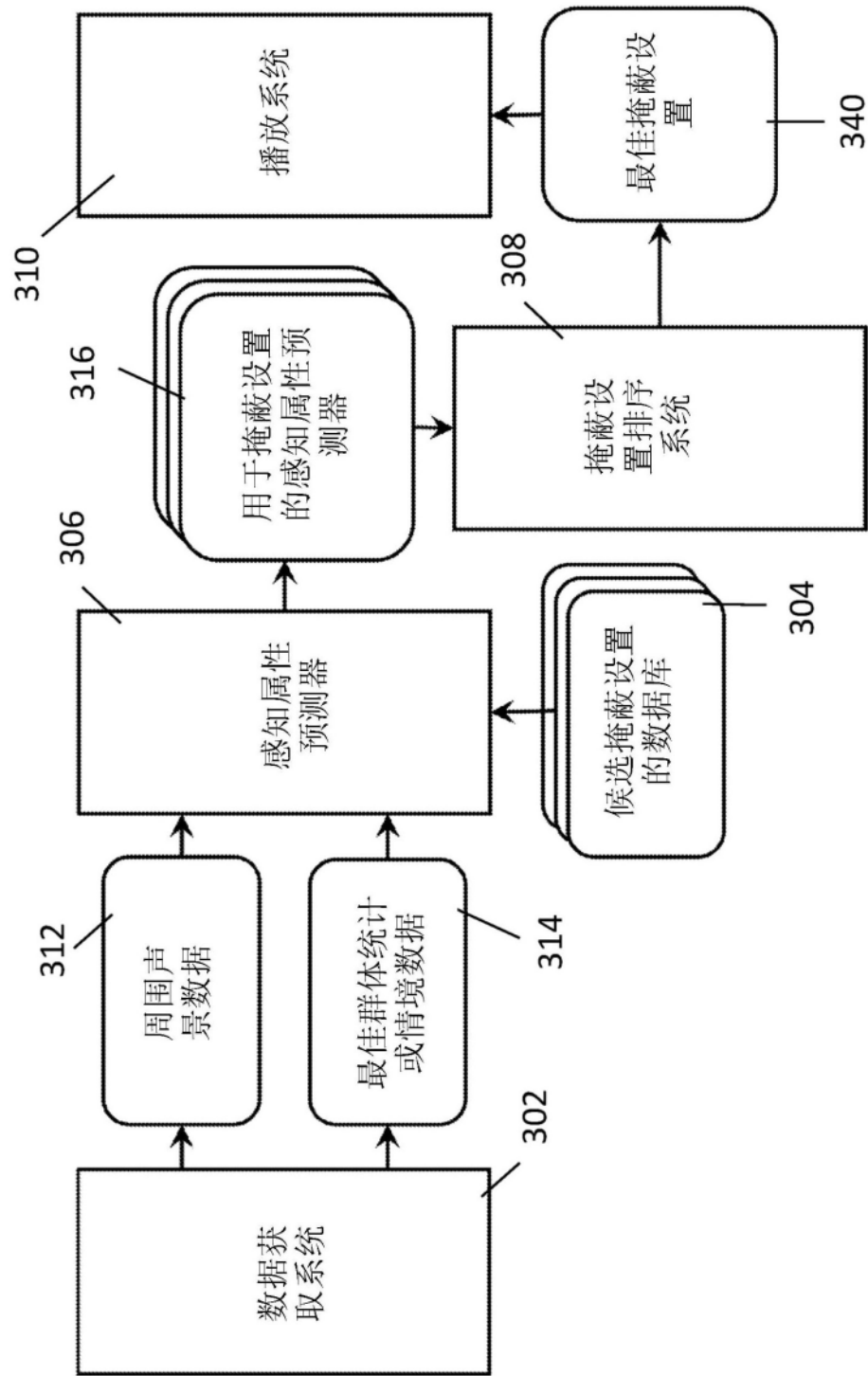


图3

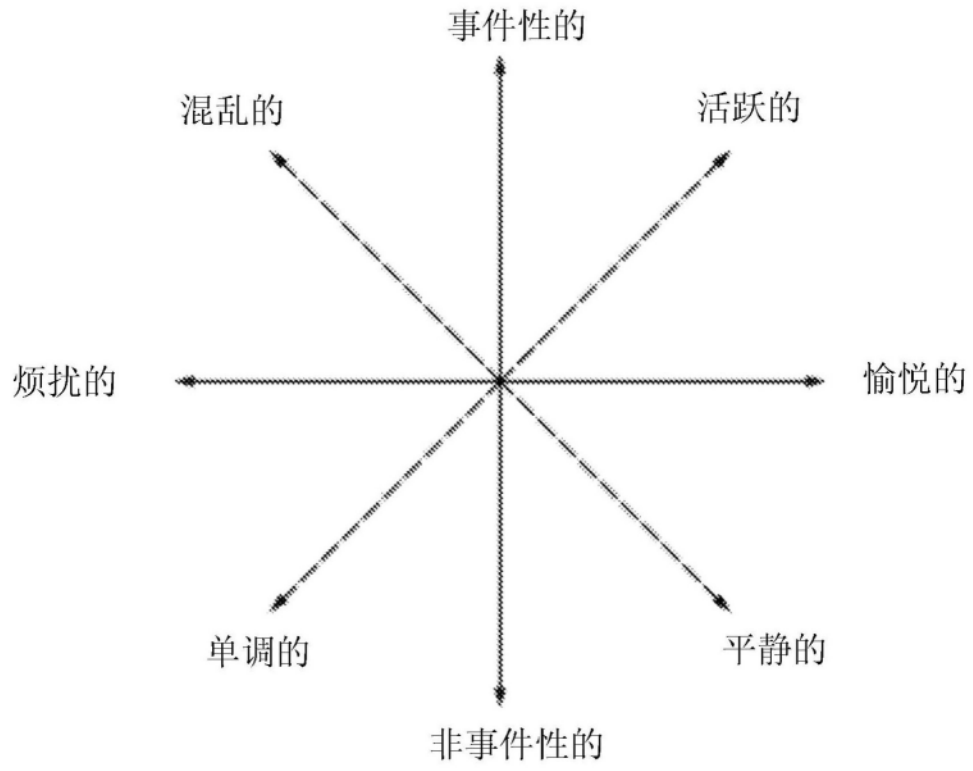


图4

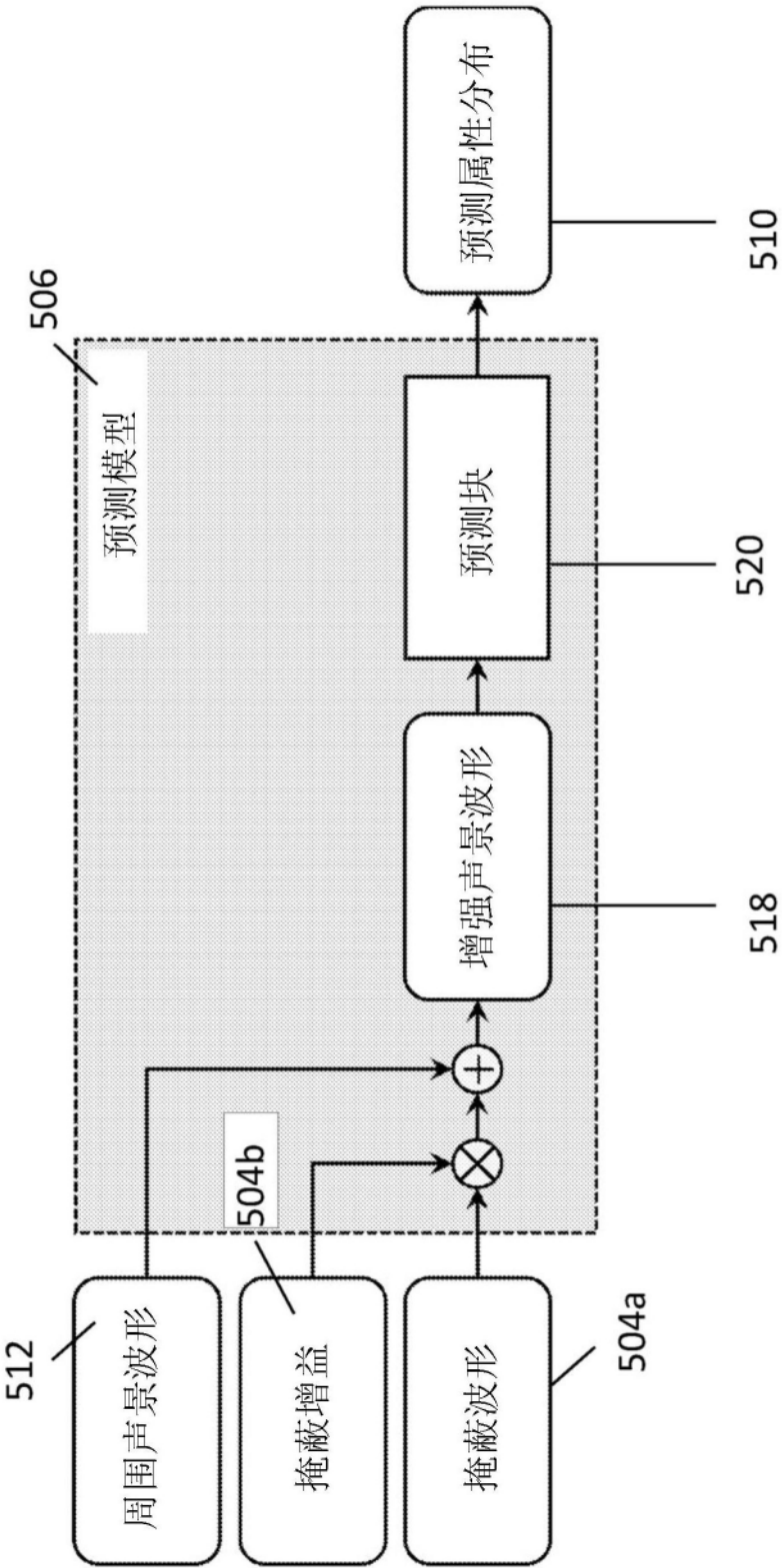


图5

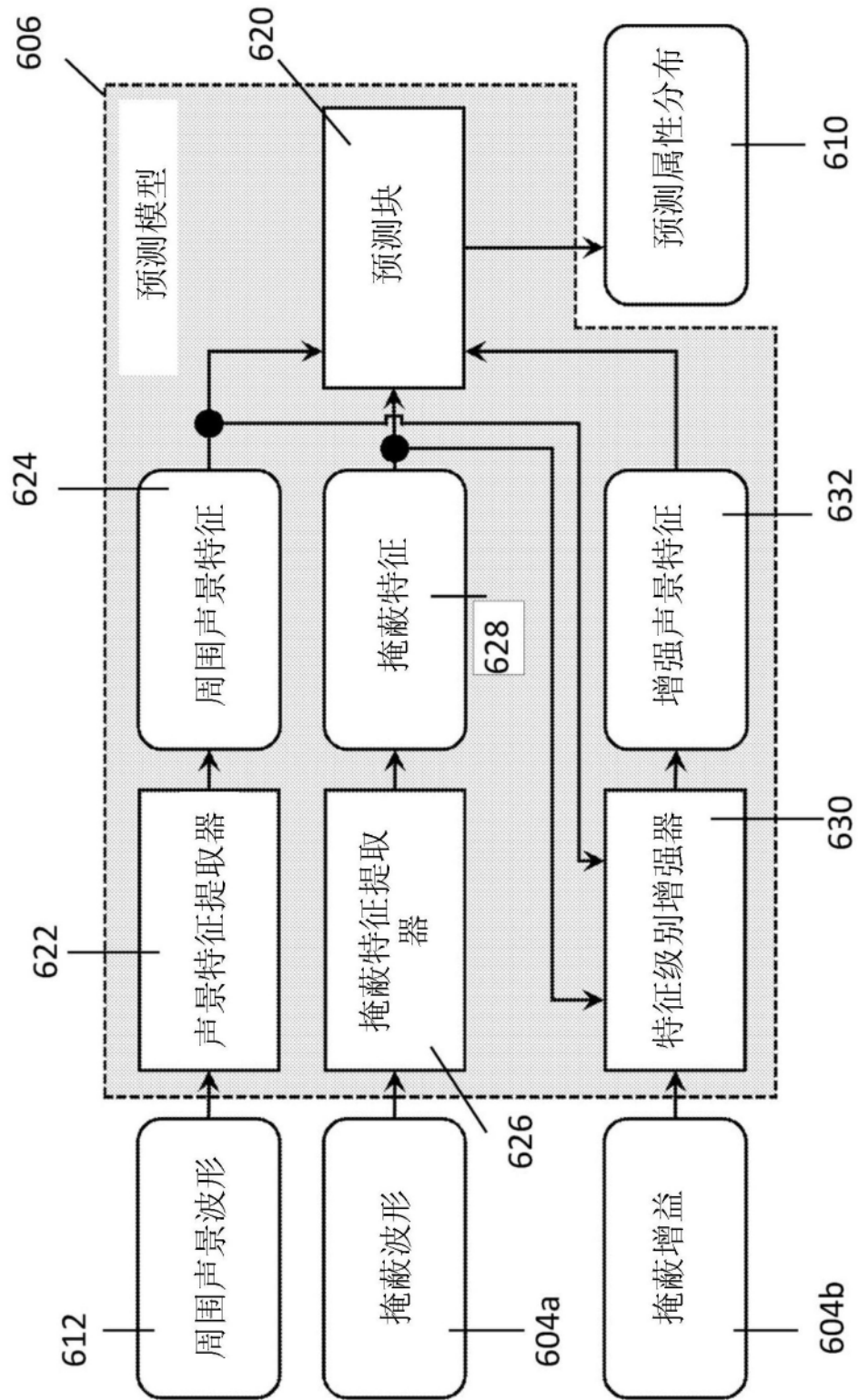


图6

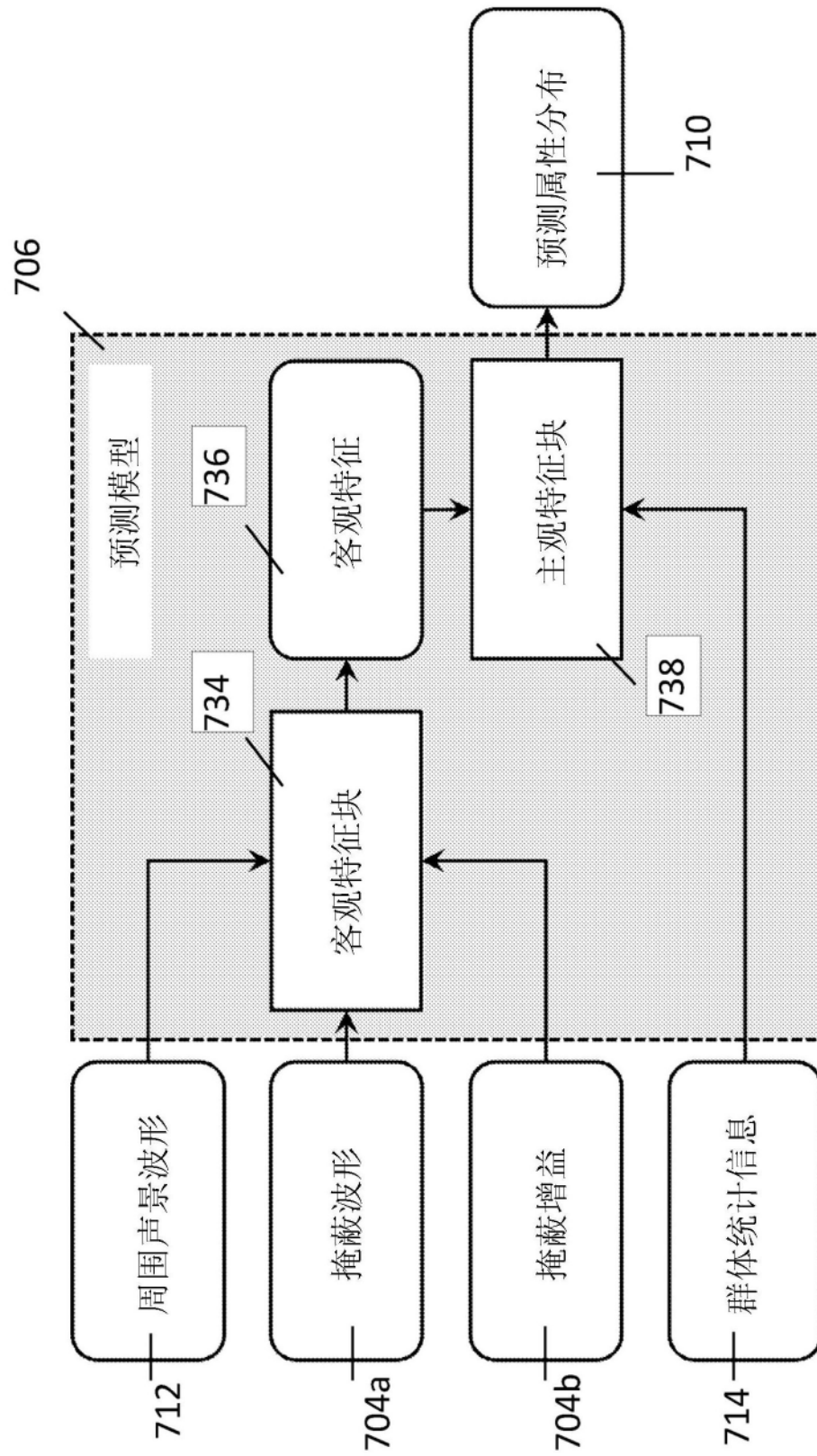


图7

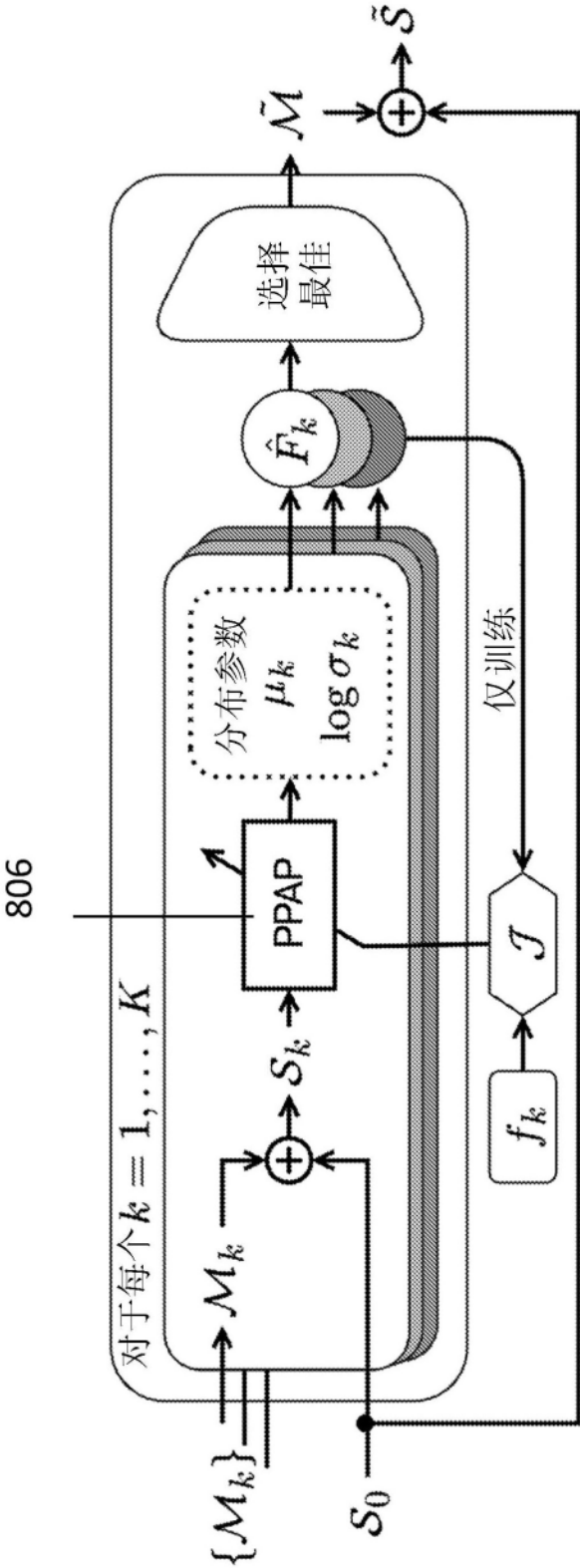


图8

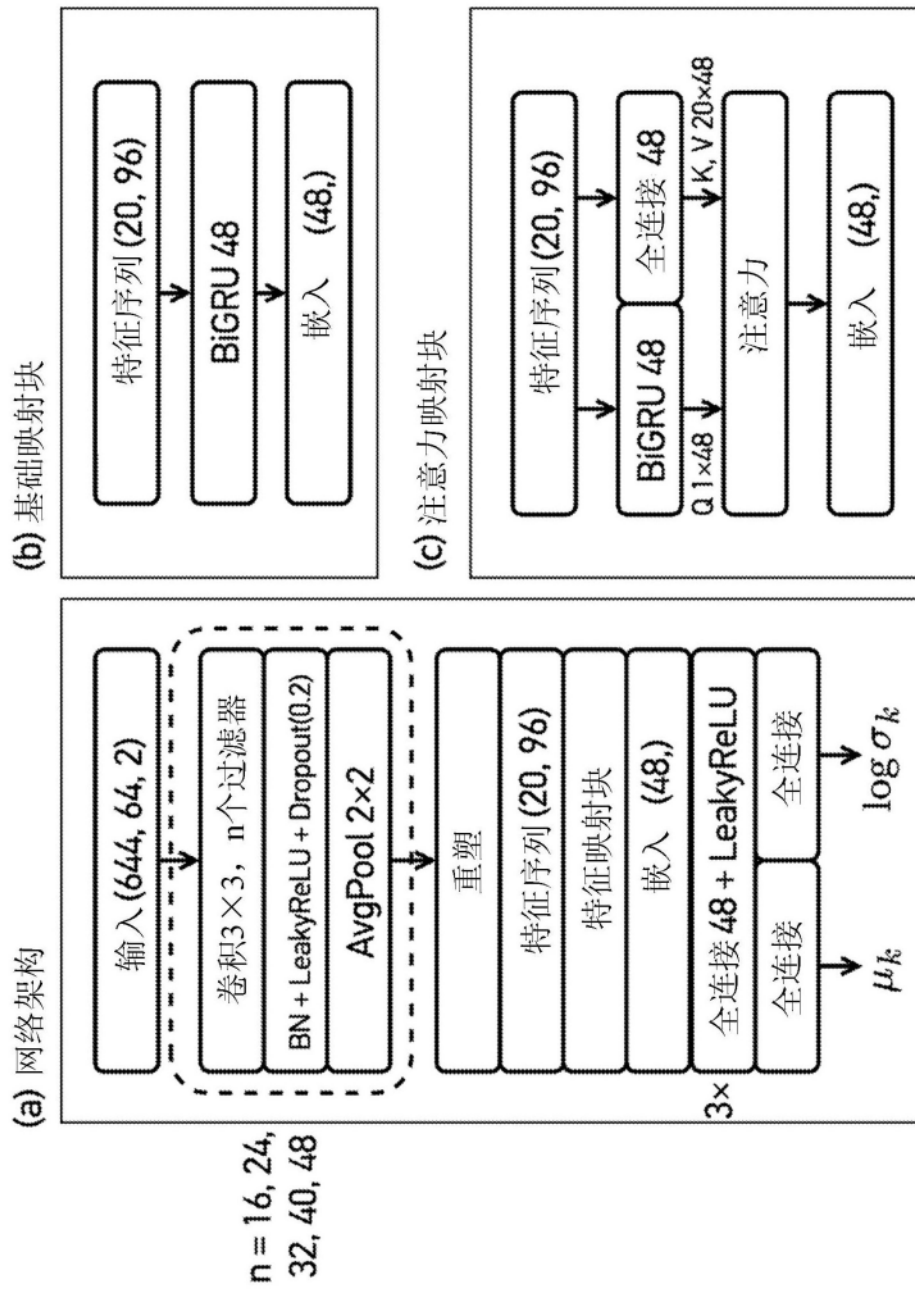


图9

交叉验证集				
模型	参数	确定性的	概率性的	改进 (%)
(标签均值)	—	0.1556	—	—
无注意力的PPAP	76K	0.1262±0.0009	0.1255±0.0011	0.5
具有加性注意力的PPAP	85K	0.1264±0.0015	0.1251±0.0014	1.0*
具有点积注意力的PPAP	85K	0.1259±0.0012	0.1250±0.0013	0.7*
具有多头注意力的PPAP	123K	0.1244±0.0009	0.1230±0.0010	1.1*

测试集				
模型	参数	确定性的	概率性的	改进 (%)
(标签均值)	—	0.1181	—	—
无注意力的PPAP	76K	0.1106±0.0114	0.1020±0.0096	7.8*
具有加性注意力的PPAP	85K	0.1030±0.0137	0.0970±0.0080	5.8*
具有点积注意力的PPAP	85K	0.0950±0.0131	0.0906±0.0091	4.6
具有多头注意力的PPAP	123K	0.0999±0.0147	0.0996±0.0104	0.3

图10

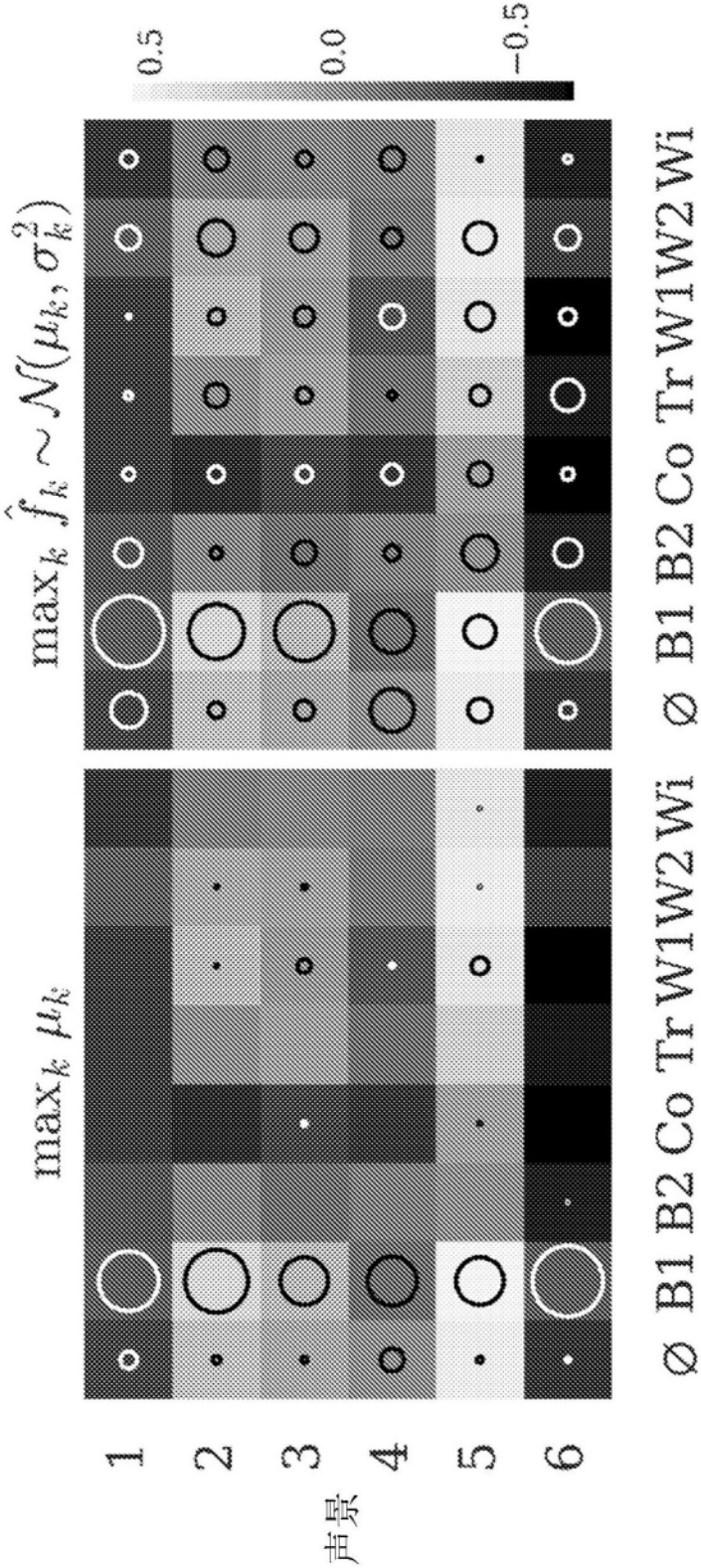


图11

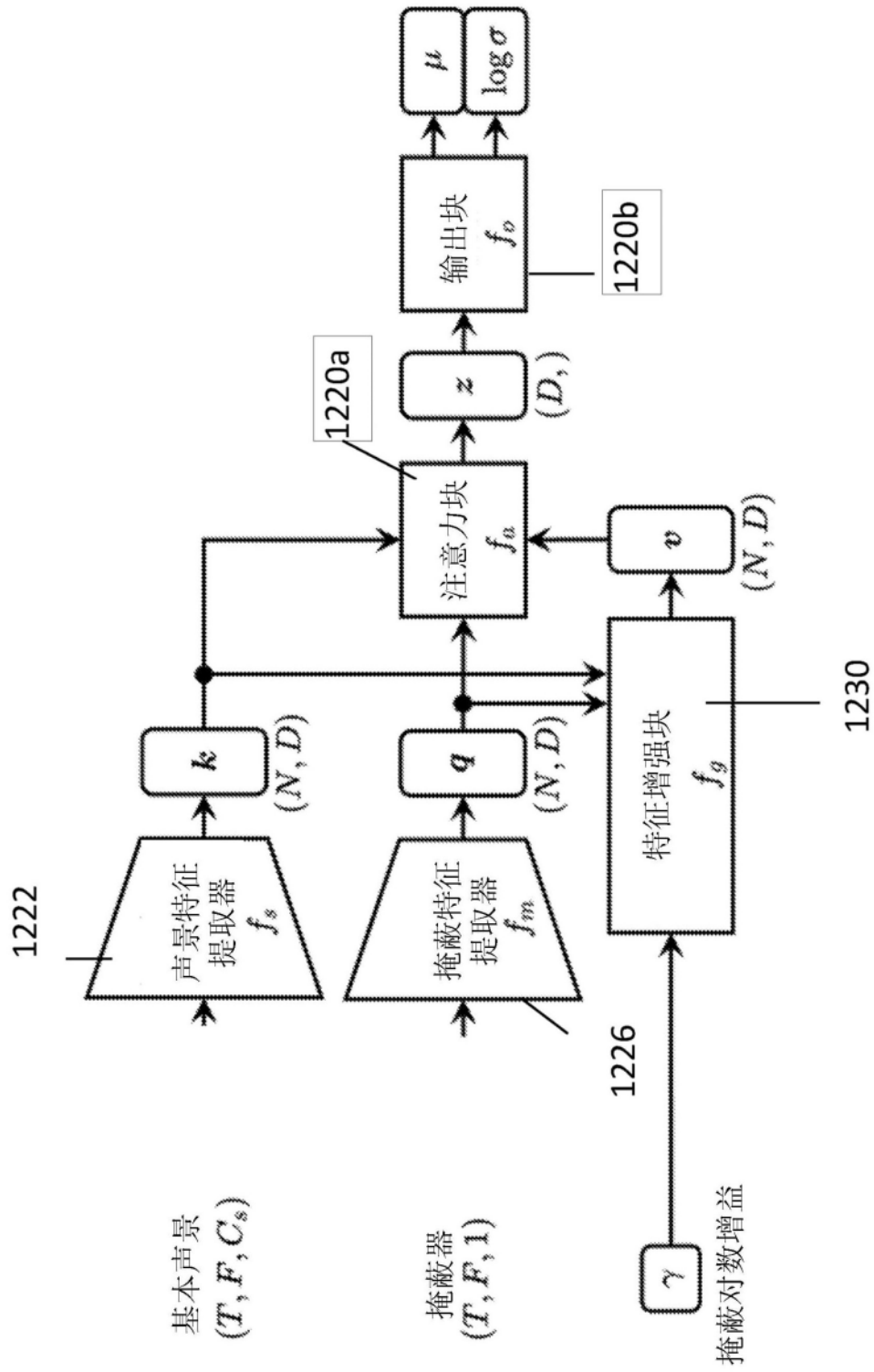


图12A

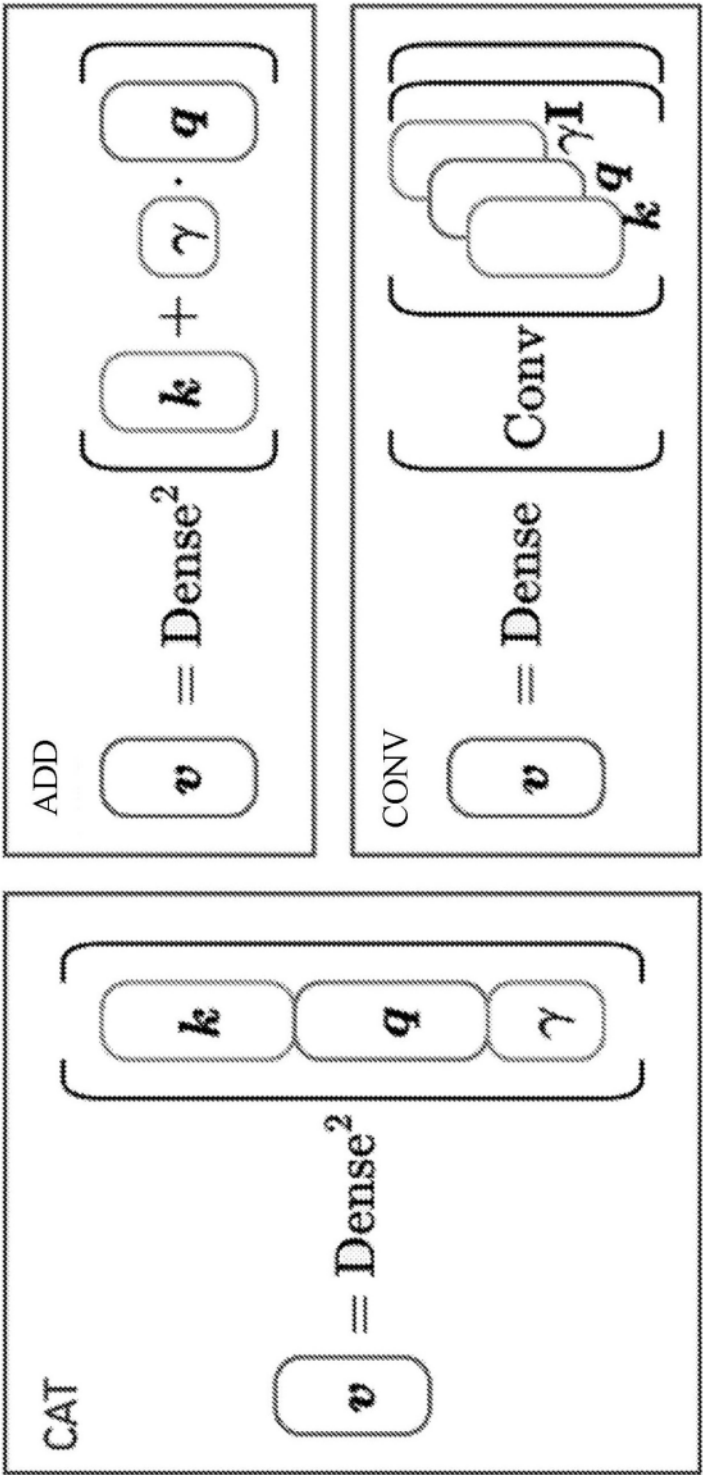


图12B

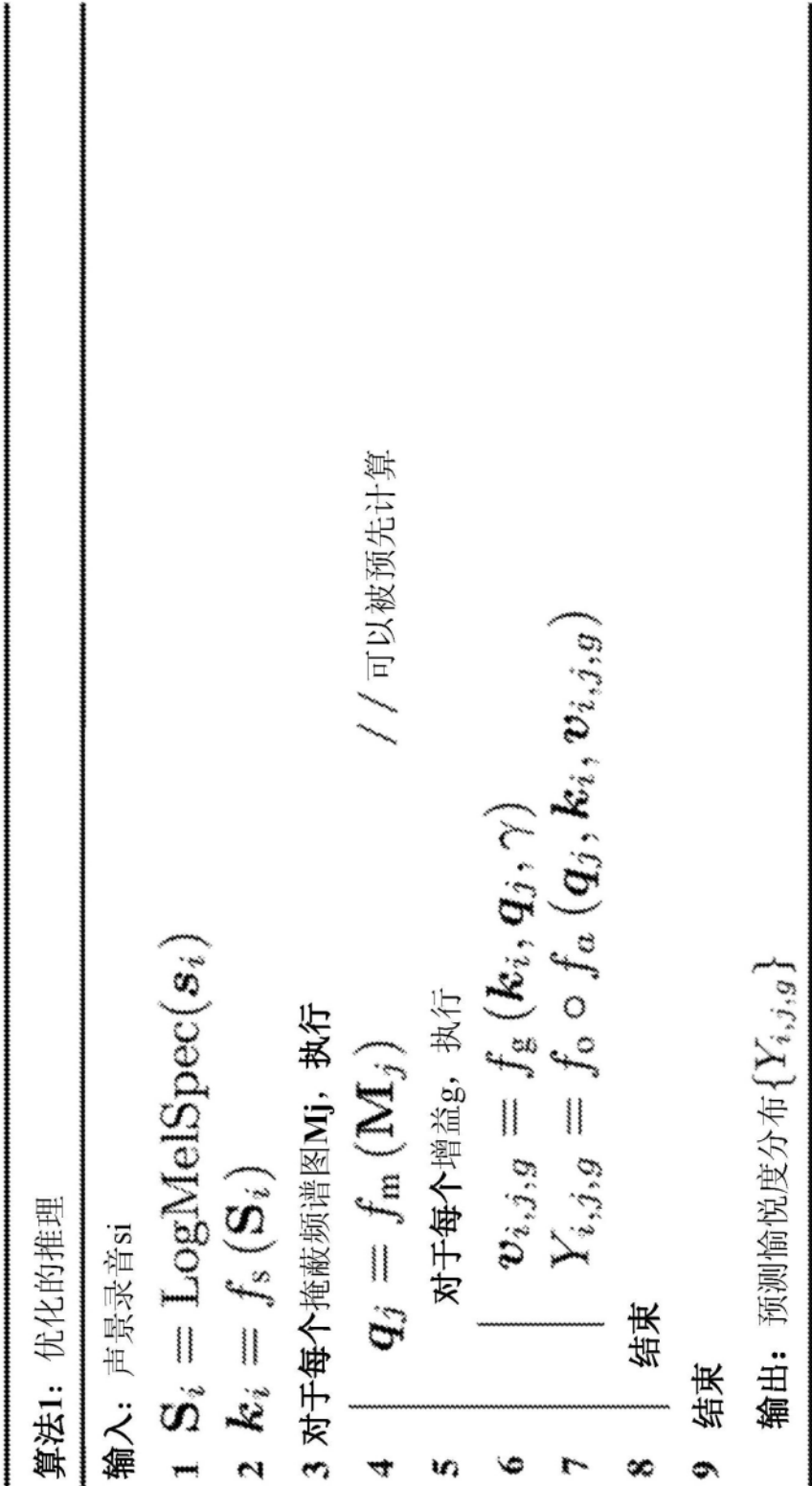


图13

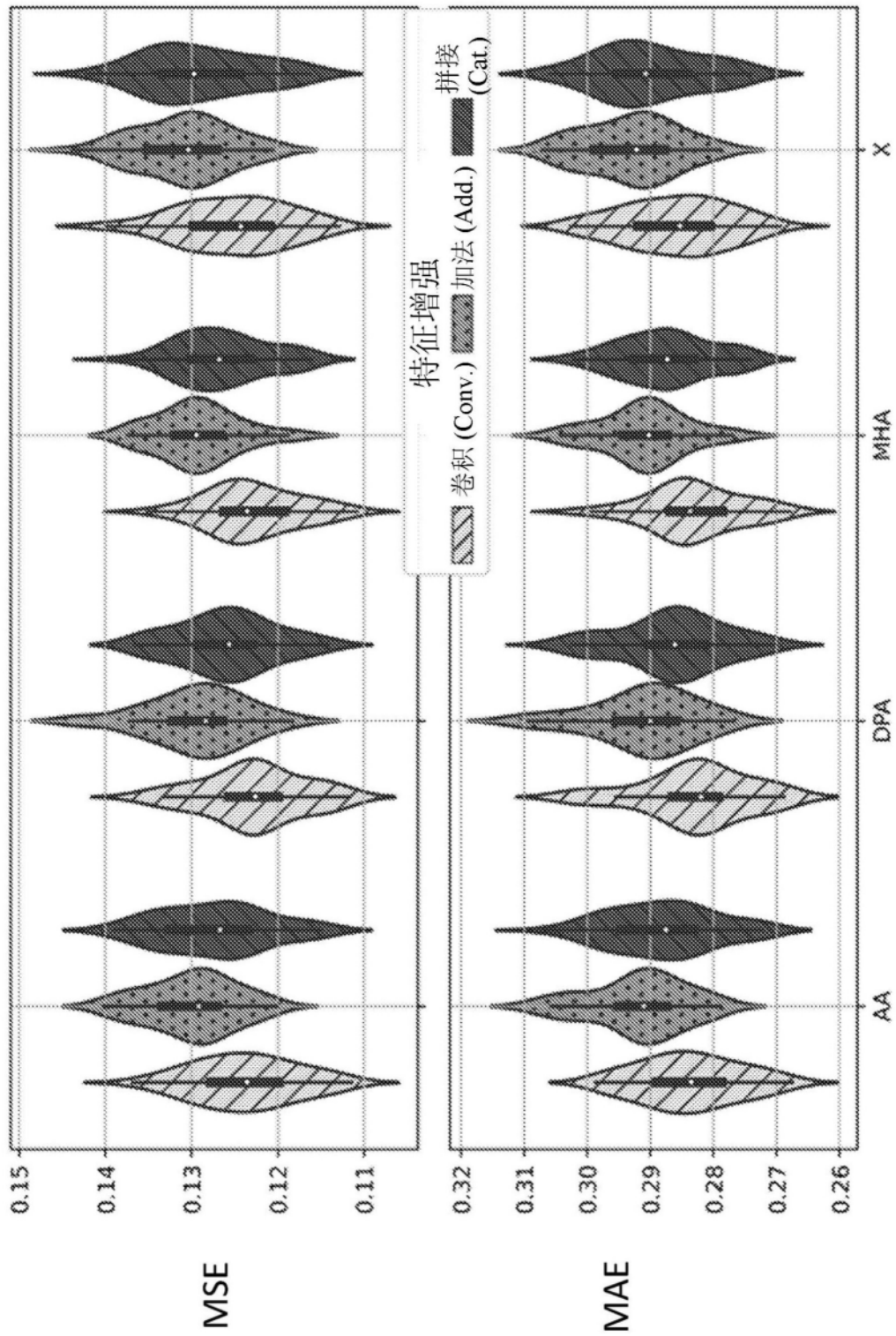


图14

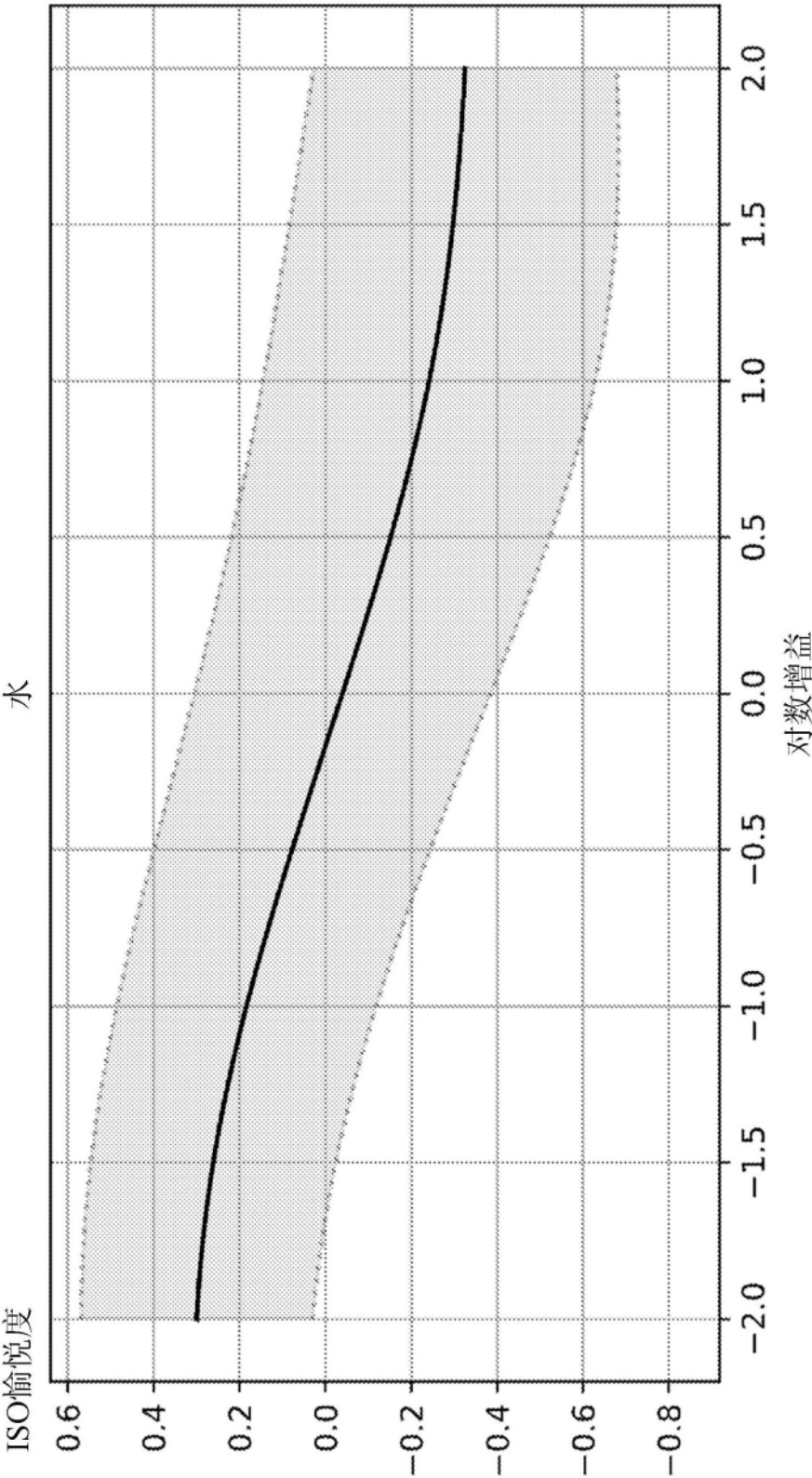


图15A

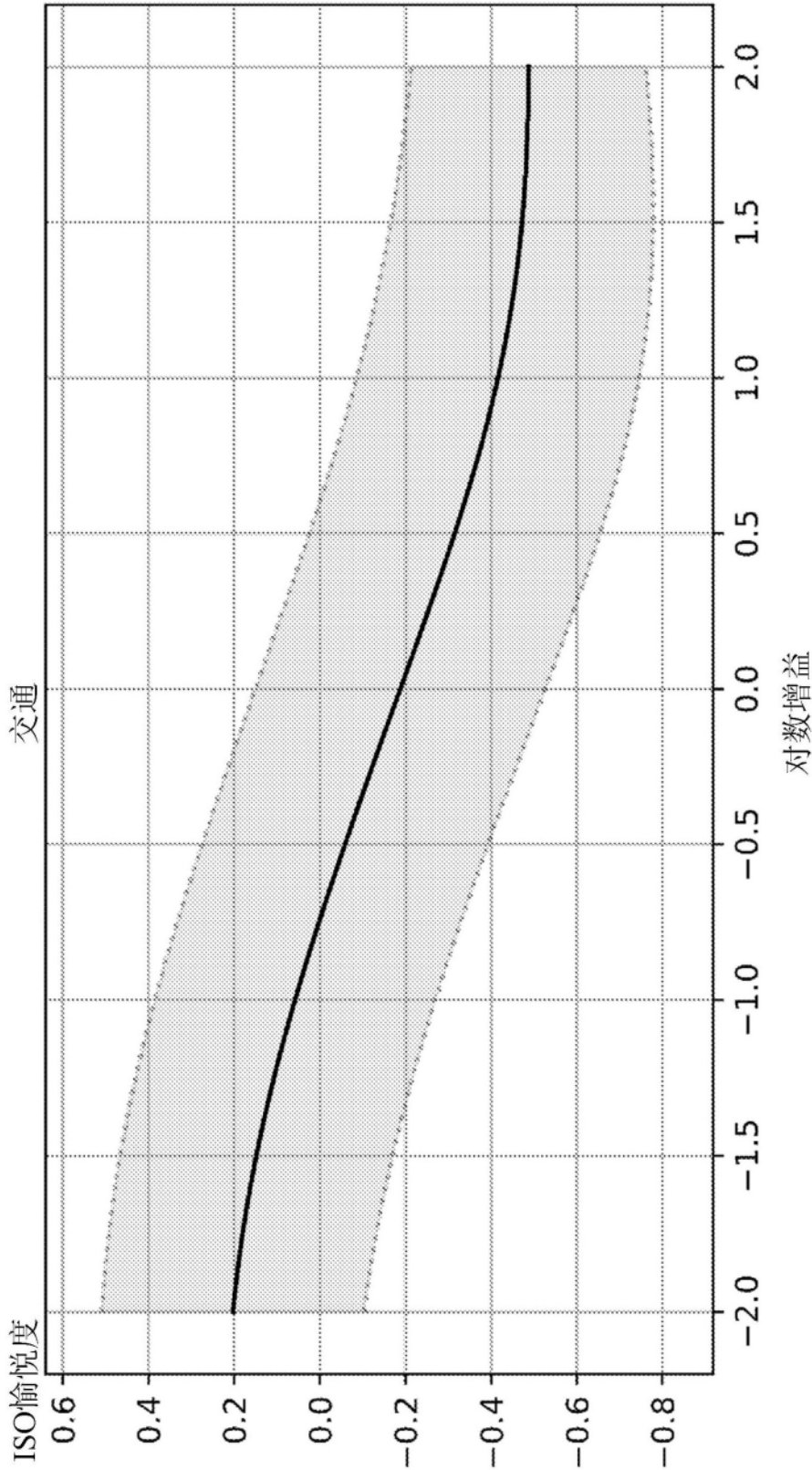


图15B

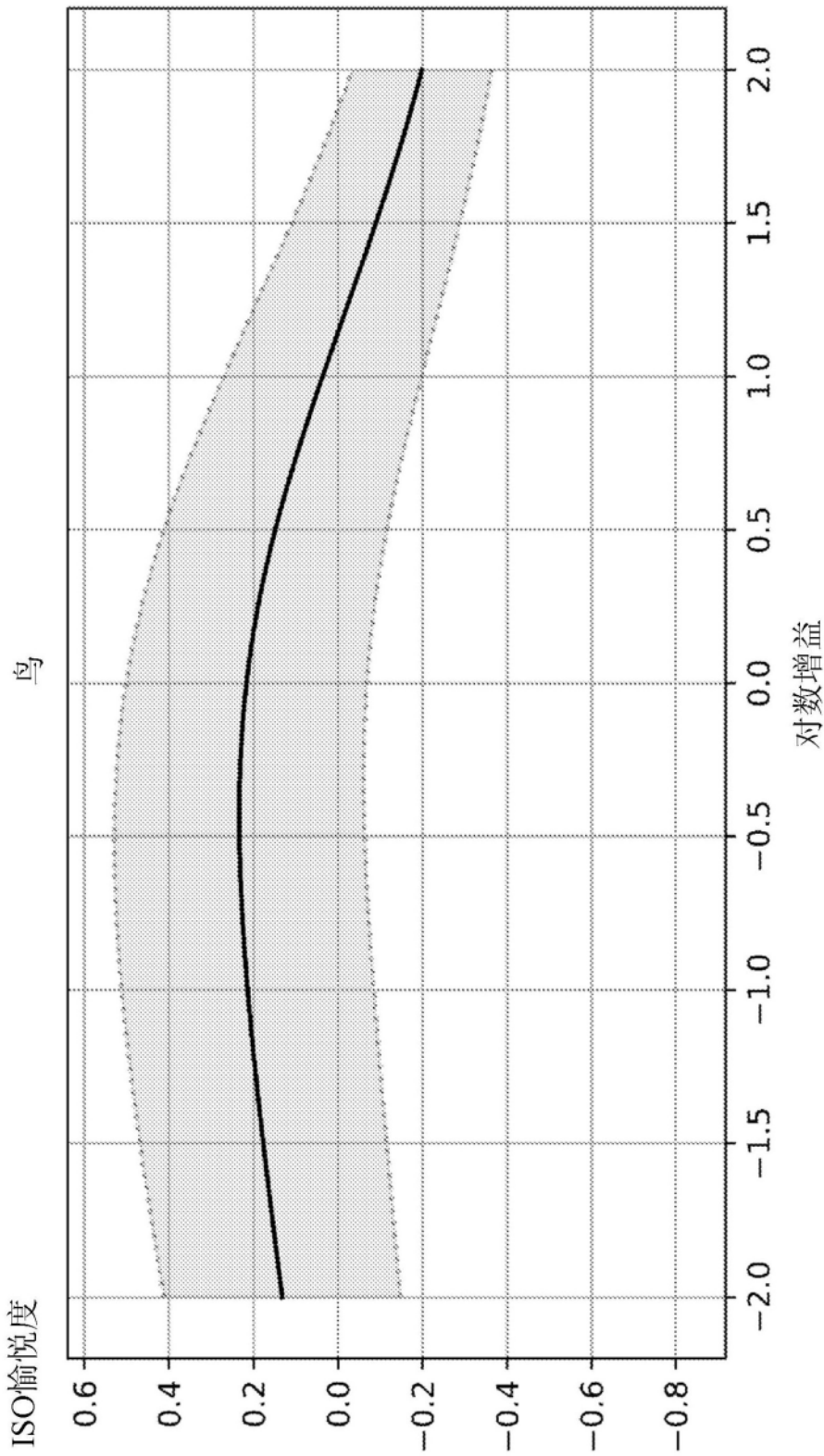


图15C

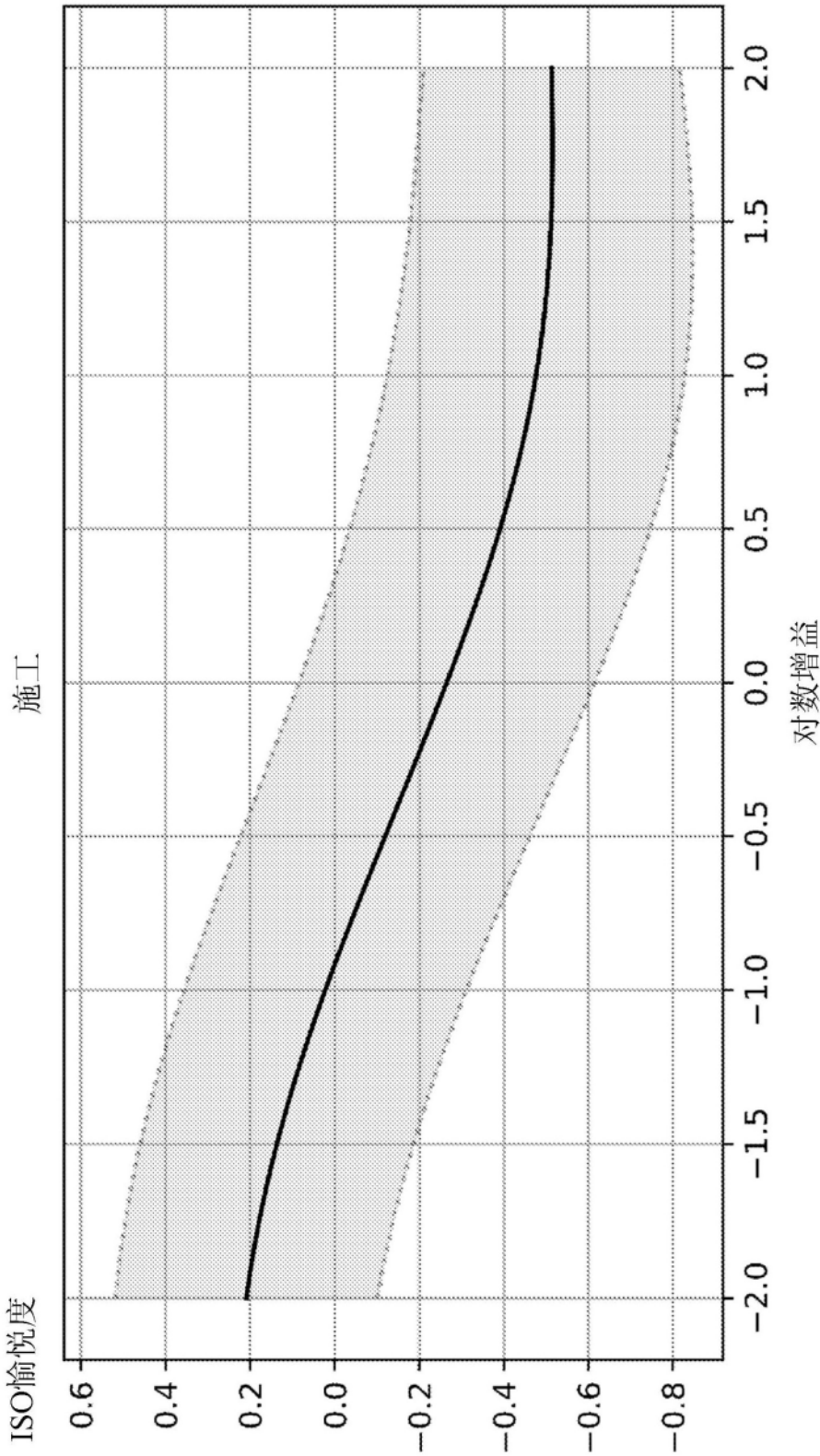


图15D

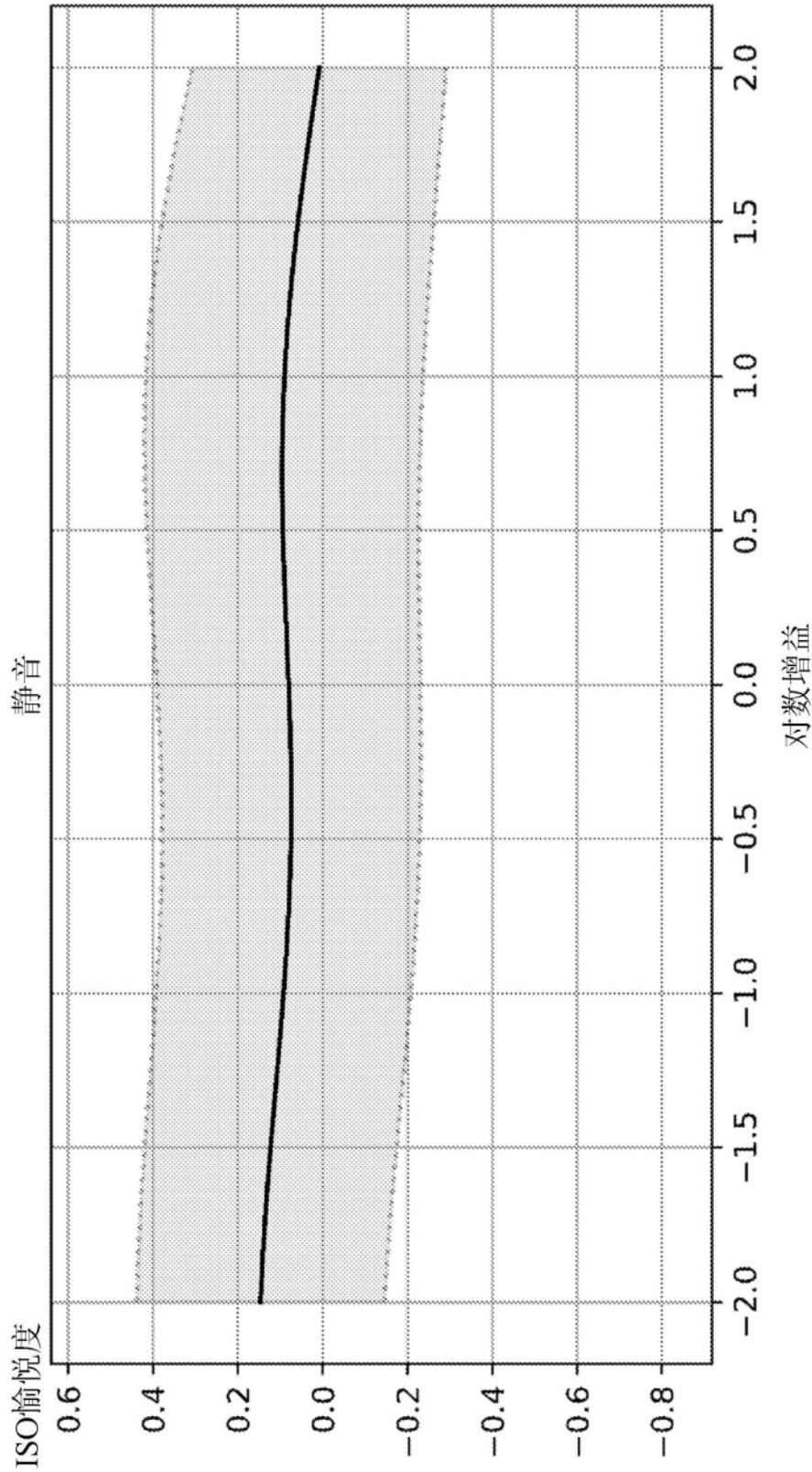


图15E