

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

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

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

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

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

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

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

where v_{pl} , 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.