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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 May 2023 (11.05.2023)

WIPO | PCT

(10) International Publication Number
WO 2023/077237 A1

(51) International Patent Classification:
H04R 3/00 (2006.01) *G10L 25/30* (2013.01)

(21) International Application Number:
PCT/CA2022/051637

(22) International Filing Date:
04 November 2022 (04.11.2022)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/275,636 04 November 2021 (04.11.2021) US

(71) Applicant: **TANDEMLAUNCH INC.** [CA/CA]; 780 Brewster Ave., Suite RC-016, Montreal, Québec H4C 2K1 (CA).

(72) Inventors: **MACCONNELL, Duncan**; c/o TandemLaunch Inc., 780 Brewster Ave., Suite RC-016, Montreal, Québec H4C 2K1 (CA). **GOLSHANARA, Ladan**; c/o TandemLaunch Inc., 780 Brewster Ave., Suite RC-016, Montreal, Québec H4C 2K1 (CA). **TURRUBIATES, Jesus**; c/o TandemLaunch Inc., 780 Brewster Ave., Suite RC-016, Montreal, Québec H4C 2K1 (CA).

(74) Agent: **WRIGHT, Laurie C.** et al.; Blake, Cassels & Graydon LLP, Commerce Court West, 199 Bay Street, Suite 4000, Toronto, Ontario M5L 1A9 (CA).

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

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*
— *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

(54) **Title:** SYSTEM AND METHOD OF IMPROVING AN AUDIO SIGNAL

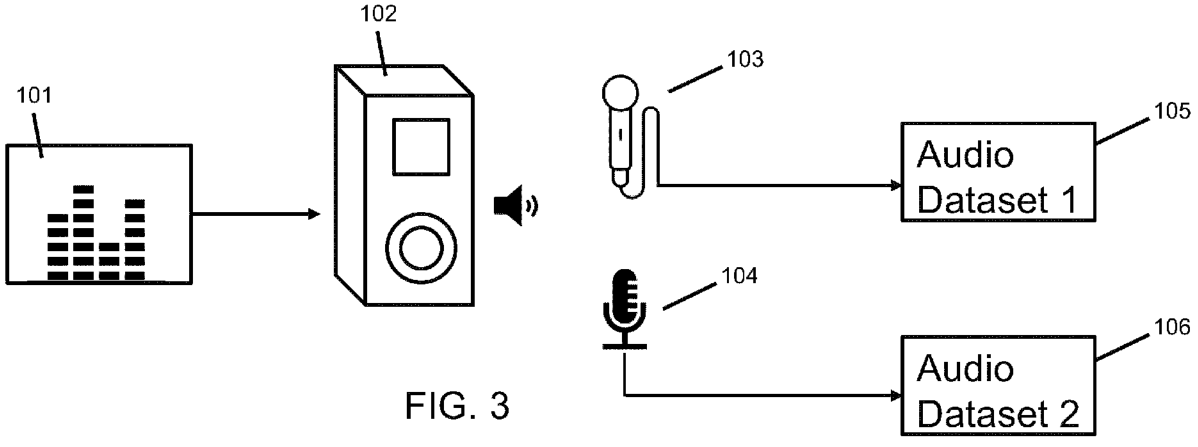


FIG. 3

(57) **Abstract:** A method of improving an audio signal is taught herein. The method comprising: outputting an audio waveform from a sound source; capturing the audio waveform from a first microphone and capturing the audio waveform from a second microphone capsule aligned beside the first microphone; and sending the captured audio waveforms to a digital audio processing system having a neural network. The neural network is configured to learn differences between the first audio waveform and the second audio waveform. The audio signals processed from the first microphone differing from the audio signal processed from second microphone. The sound source may comprise a curated data set consisting of test signals, representative audio signals.

SYSTEM AND METHOD OF IMPROVING AN AUDIO SIGNAL

FIELD OF THE DESCRIPTION

[100] The following relates to audio signals generated from audio devices such as microphones and speakers. The following more specifically relates to the improvement and processing of these audio signals.

BACKGROUND

[200] FIG. 1 provides a schematic diagram of a conventional condenser microphone. Condenser microphones work on the principle of capacitance. Capacitors consist of parallel conducting plates that store charge and are used to smooth out signals like voltage variations in a power supply. In a condenser microphone, the incoming sound 1 vibrates the diaphragm 2 of a capacitor. This varies the capacitance between the diaphragm 2 and the back plate 3. The varying capacitance is converted into a corresponding electrical signal 4.

[300] FIG. 2 provides a schematic diagram of a conventional dynamic microphone. A dynamic microphone converts sound into a small electrical current. Sound waves 1 hit a diaphragm 2 that vibrates, moving a magnet 8 near a coil 7. This produces an electric current 9.

[400] Both the condenser and dynamic microphones are transducers; they transform sound pressure waves into voltage, through the movement of the microphone diaphragm 2. The selection of diaphragm 2 and accompanying electrical circuit determine the voltage that represents the sound, and thus determine the perceptual quality of the sound. It is very expensive to design and build high quality, professional microphones. When product designers need microphones in products, the size and expense of the microphone is weighed against the perceptual benefits of having a high-quality microphone in the system.

[500] One factor affecting the quality of the microphone is the build quality of the diaphragm 2, The sensitivity of the diaphragm is dependent on the size of the diaphragm 2. Diaphragm material, design, thickness, and diameter can help to determine a microphone’s frequency, transient and polar responsiveness. In turn, the microphone quality is limited by, for example, the material, design, thickness, and diameter of the Diaphragm 2.

[600] Typically, diaphragms 2 can be categorized into three sizes—large, medium, and small. Larger diaphragm microphones are typically more sensitive due to their increased surface area, but also have a more limited frequency response since sound waves have to move more mass.

[700] Small diaphragm microphones are capable of handling higher sound pressure levels due to their stiffer diaphragms. They also have an increased frequency response, particularly in the higher end of the frequency spectrum. Their decreased sensitivity relative to large diaphragm microphones makes them less susceptible to proximity effect and ambient noise due to their directional characteristics.

[800] As such, manufacturers of phones and other devices which use microphones are required to choose between a lower cost per unit or a higher quality of sound. Significant efforts have been made to develop lower cost microphones, that can achieve high level of sound.

[900] Furthermore, the characteristics of various microphones make them more suitable for certain applications. For instance, condenser microphones are better suited for high frequency applications such as recording a vocalist in an isolation booth, recording an acoustic guitar to capture definition, recording a group of singers, recording an acoustic piano, recording sound effects, or recording a podcast voice in a quiet or acoustically treated room.

[1000] On the other hand, dynamic microphones are not as sensitive, which makes them better suited for low frequency applications such as recording drums, recording guitar amplifiers, recording multiple individuals' voices sitting around a table, or recording one or more speakers on a stage when you need to avoid picking up other sounds.

[1100] As such, the same microphone cannot be used for a variety of applications. This can lead to microphone users having to purchase many different microphones to record different kinds of sounds, which can be expensive.

[1200] Improved systems are needed.

SUMMARY OF THE DESCRIPTION

[1300] In one aspect, there is provided a method of improving an audio signal comprising: outputting an audio waveform from a sound source; capturing the audio waveform from a first microphone and capturing the audio waveform from a second microphone capsule aligned beside the first microphone; and sending the captured audio waveforms to a digital audio processing system having a neural network. The neural network is configured to learn differences between the first audio waveform and the second audio waveform. The audio signals processed from the first microphone differing from the audio signal processed from second microphone. The sound source may comprise a curated data set consisting of test signals, representative audio signals.

[1400] The method may further comprise applying the learned differences to a third audio waveform recorded from a third microphone; such that the third microphone has similar characteristics to the first microphone.

[1500] The first microphone may have a non-ideal set of characteristics, and the second microphone may have an ideal set of characteristics for a specific function. The specific function can be selected from at least one of the following: conversation, lyrical, music, noise cancellation, and instrumental.

[1600] The third microphone can be located on a mobile device. The third microphone records audio from a telephone conversation such that the digital audio processing system processes the conversation in real-time. The digital audio processing system can be located on an application on the mobile device.

[1700] The method can also be used for polar pattern translation, wherein the first microphone comprises a first polar pattern from one of a unidirectional, bidirectional, and omnidirectional patterns, and the second microphone comprises a second polar pattern, from one of a unidirectional, bidirectional, and omnidirectional microphones. The third microphone may comprise the first polar pattern and wherein the learned differences between the first and second microphone can be applied to a third audio waveform recorded from the third microphone to match the second polar pattern.

BRIEF DESCRIPTION OF THE FIGURES

[1800] The features of certain embodiments will become more apparent in the following detailed description in which reference is made to the appended figures wherein:

[1900] FIG. 1 depicts a schematic diagram of a condenser microphone;

[2000] FIG. 2 depicts a schematic diagram of a dynamic microphone;

[2100] FIG. 3 depicts a schematic diagram of recording a training model dataset;

[2200] FIG. 4 depicts a schematic diagram of the model training inputs;

[2300] FIG. 5 depicts a schematic diagram of the digital audio processing system algorithm;

[2400] FIG. 6 depicts an embodiment of the application of the invention;

[2500] FIG. 7 depicts a further embodiment of the application of the invention;

[2600] FIG. 8A depicts a schematic diagram of a unidirectional polar pattern;

[2700] FIG. 8B depicts a schematic diagram of a bi-directional polar pattern;

[2800] FIG. 8C depicts a schematic diagram of an omnidirectional polar pattern;

[2900] FIG. 9A depicts a schematic diagram of the digital audio processing system;

[3000] FIG. 9B depicts a schematic diagram of the digital audio processing system;

[3100] FIG. 10 depicts a schematic diagram of the deep neural network algorithm;

[3200] FIG. 11 depicts a schematic diagram of the digital audio processing system algorithm for upgrading a speaker;

[3300] FIG. 12A depicts one example of the training model; and

[3400] FIG. 12B lists the equations used by the training model.

DETAILED DESCRIPTION

[3500] The terms “comprise”, “comprises”, “comprised” or “comprising” may be used in the present description. As used herein (including the specification and/or the claims), these terms are to be interpreted as specifying the presence of the stated features, integers, steps, or components, but not as precluding the presence of one or more other feature, integer, step, component, or a group thereof as would be apparent to persons having ordinary skill in the relevant art. Thus, the term "comprising" as used in this specification means "consisting at least in part of. When interpreting statements in this specification that include that term, the features, prefaced by that term in each statement, all need to be present but other features can also be present. Related terms such as "comprise" and "comprised" are to be interpreted in the same manner.

[3600] Unless stated otherwise herein, the article “a” when used to identify any element is not intended to constitute a limitation of just one and will, instead, be understood to mean “at least one” or “one or more”.

[3700] The invention provides a Digital Audio Processing System 100 having a neural net audio processing translation from one distinct microphone to another microphone. This invention allows small and cheap microphones to take on the spectral characteristics (i.e. the particular sound) of a large high-quality microphone.

[3800] The Digital Audio Processing System 100 uses a neural network to learn the difference between audio signals captured by model microphones and audio signals captured by lower-quality microphones. The system then applies the trained model differences on the audio signals of other lower-quality microphones to produce high-quality sounds from the lower-quality microphones. Machine learning allows us to learn the complex effect that microphones have when capturing sound.

[3900] FIG. 3 depicts a schematic diagram of recording a training model dataset. An unprocessed audio dataset 101 is played on a speaker 102. The unprocessed audio dataset 101 may be any type of unprocessed audio such as a recording of a group of people having a conversation, a recording of a musical instrument, a recording of a person or group of people singing, musical songs, etc. In another embodiment, the unprocessed audio dataset may be any source of sound, such as a group of people having a conversation, a musical instrument, a person, or group of people singing, etc.

[4000] At the same time, two microphones 103, 104 are used to capture and record the unprocessed audio signal 101 simultaneously. The two microphones 103, 104 are preferably capsule aligned such that they are capturing an identical audio wavefront. Capsule aligned refers to the physical alignment of the diaphragm 2 of two or more microphones.

[4100] In one embodiment, one of the microphones 103 may be of lesser quality than the other microphone 104. The build quality of the microphone will affect the sound quality recorded by the microphone. For instance, diaphragm material, design, thickness, and diameter can help to determine a microphone's frequency, transient and polar responsiveness. In turn, the microphone quality is limited by, for example, the material, design, thickness, and diameter of the diaphragm. For instance, microphone 103 may be a simple, inexpensive microphone and microphone 104 can be a high-quality recording microphone which is expensive.

[4200] In this embodiment, since the first microphone 103 is of lesser quality than the second microphone 104, the resulting audio recordings or datasets 105, 106 will be of varying quality. The first dataset 105 is the audio recorded by the first microphone 103. The second dataset 106 is the audio recorded by the second microphone 104. In this embodiment, we assume the second microphone 104 is of higher quality, and thus the recorded audio signal 106 will have a higher sound quality (i.e., less distortion, less static noise, high bit rate, etc.) than audio signal 105. Consequently, we assume that if the first microphone 103 is of lesser quality, the audio signal 105 recorded by the first microphone 103 will have a lower sound quality compared to the audio signal 106 recorded by the second microphone 104. As such, audio signal 105 is expected to have more distortion, more static noise, low bit rate, than audio signal 106 as audio signal 105 is recorded from a low-quality microphone 103.

[4300] In another embodiment, the first microphone 103 can have a first set of characteristics and the second microphone 104 can have a second set of characteristics. These characteristics may include differences in the microphone's capturing frequency, quality, type of sound, etc. In a third embodiment, the first microphone 103 may be a condenser-type of microphone and the second microphone 104 may be a dynamic-type of microphone. As such, the resulting dataset 105 recorded by the first microphone 103 will be different from the dataset 106 recorded by the second microphone 104.

[4400] The varying recorded audio signals 105 and 106 is then input into a deep neural network (DNN) 107. The difference between the audio signals 105, 106 of the two different microphones 103, 104 is captured, the neural network 107 is configured to learn the difference between the microphones 103, 104. The Digital Audio Processing System 100 can then apply this difference to new audio signals. FIG. 4 shows a schematic diagram of the model training inputs to be used by the deep neural network 107 and Digital Audio Processing System 100.

[4500] Many different high quality microphones can be used to train the neural network 107 since the characteristics of different microphones make them more suitable for certain applications. For instance, condenser microphones are better suited for high frequency applications such as recording a vocalist in an isolation booth, recording an acoustic guitar to capture definition, recording a group of singers, recording an acoustic piano, recording sound effects, or recording a podcast voice in a quiet or acoustically treated room. On the other hand, dynamic microphones are better suited for low frequency applications such as recording drums, recording guitar amplifiers, recording multiple individuals' voices sitting around a table, or recording one or more speakers on a stage when you need to avoid picking up other sounds. As such, varying types of microphones can be used to train the data models for varying applications. In another embodiment, a plurality of high quality microphones can be used simultaneously to train the model.

[4600] The audio dataset 105 processed by the first microphone 103 and the audio dataset 106 processed by the second microphone 104, once recorded through an analog to digital converter and into a digital audio workstation 113, can be used as digital audio inputs to the deep neural network 107. The neural network 107 can use distinct digital audio inputs 105 and 106 and learn the differences between the wave forms.

[4700] FIG. 5 depicts a schematic diagram of the digital audio processing system algorithm for upgrading a microphone. A microphone use-case model, or representative model is to be determined 501. This can be done via a computational processor or as a user input. The representative models include (but are not limited to): ameliorating microphone quality, polar pattern translation, microphone position modelling, noise cancellation, singing, conversational audio, lyrical, instrumental audio, etc. Audio data from the use case is then obtained 502. At least a first and second microphone 103, 104 are then obtained 503. These microphones can be differing in some way such as quality, polar pattern, position, etc.). Any number of microphones can be used to train the data. Subsequently, a form of audio is played on a speaker and recorded on microphones 103, 104. The form audio is ideally related to the use case (i.e. conversation, music, lyrical, noise cancellation, instrumental, etc.). It is also ideal if the diaphragms of the microphones 103, 104 are aligned. The audio signals 105 and 106 can be obtained by simultaneously recording the audio on microphones 103 and 104 (see 504 of FIG. 5). As a result, at least two datasets 105, 106 are obtained from the recording captured by microphones 103, 104 (see 505 and 506 of FIG. 5). The audio dataset 505 and 506 can then be used as an input for a deep neural network 107 (see 507 of FIG. 5). The neural network outputs a set of learned weights for the deep neural network layers (508). Neural network architecture is implemented to learn processing (509). The learned weights, or learned differences, can then be applied to new audio that has not been seen by the DNN (510). The new audio may be in the form of a pre-recorded digital signal or on real-time audio converted to a digital signal. As a result, the new audio, which may have been recorded on a microphone of lesser quality, or microphone of non-ideal characteristics, may be upgraded to sound like audio recorded on a high quality microphone, or a microphone of ideal characteristics.

[4800] The deep neural network 107 training model expects to receive representative audio (502), consisting of the types of signals the in order to perform the transformation. For instance, a microphone upgrade model for a music application would require singing voice and live instrument training data, while a microphone upgrade for meeting applications would require speech signals.

[4900] Sonic characteristics include time and frequency domain characteristics, frequencies, and signal amplitudes. Various test signals comprising of singular tones, and complex combinations, at particular frequencies, for particular durations can be used as input audio signals.

[5000] This system can be used for digital emulation of specific microphone models used for recording music. In one instance, the system may be used to upgrade a budget microphone to a top-of-the-line microphone. The invention includes a neural net trained on different diaphragm aligned microphones, and the real time audio processing allowing the translation of a signal recorded with a first, lower quality microphones to sound like it was recorded with a different, higher quality microphone.

[5100] FIG. 6 depicts an embodiment of the application of the invention. In this embodiment, sound waves are captured by a new microphone 108. Microphone 108 may be of lesser quality or have a different set of characteristics than an ideal microphone 104. Microphone 108 ideally has similar characteristics as the microphone 103 that was used to train the model. The audio 110 captured by microphone 108 can be imported to the digital audio processing system 100 taught by this invention. The Digital Audio Processing System (DAPS) 100 would have been trained using the model of a high quality or, ideal microphone 104. The model trained by the ideal microphone 104 can then be applied to the signal of the new microphone 108. This ameliorates the audio 110 recorded by the new microphone 108 as the model trained by ideal microphone 104 is applied to the audio 110 of new microphone 108. The model is first trained by learning the difference between a less-than-ideal microphone 103 and an ideal microphone 104. The output of the DAPS 100 is audio 111 which sounds like it came from an ideal microphone.

[5200] FIG. 7 depicts a further embodiment of the application of the invention. In this embodiment, sound waves 110 can be recorded with a small (or low quality) microphone 701 located within a mobile device. The mobile device may include an application or program with the Digital Audio Processing System 100 on it. The Digital Audio Processing System comprises a deep neural network which can take the sound waves captured from the low-quality microphone 701 and apply the trained model to those waves. The resulting sound would be sound waves 111 having the characteristics of a higher quality microphone. As such, this process can be used by cell phone manufacturers to process Micro-Electro-Mechanical Systems (MEMS) microphone 701 signals to sound like a signal coming from a higher quality microphone. The invention includes a neural net trained on audio recorded at different microphone positions, and the real time audio processing to allow signals recorded on small MEMS microphones 701 commonly found in mobile devices, to sound like a signal that was recorded with a high-quality large diaphragm microphone. The mobile device may include an application or program with the Digital Audio Processing System 100 on it which would allow real-time phone conversations to be upgraded in real-time. In another embodiment, the DAPS 100 may be located on an external server such as a database or cloud server, wherein the input audio 110 can be exported, converted, and imported back to the mobile device. It can be understood that this model can be applied to other methods of verbal communication such as: real time audio/video calls (facetime, zoom, teams, skype, etc.), live audio streaming, live video recording/streaming, and the like.

[5300] In another embodiment, this system can be used for polar pattern translation. FIG. 8A depicts a schematic diagram of a unidirectional polar pattern; FIG. 8B depicts a schematic diagram of a bi-directional polar pattern; and FIG. 8C depicts a schematic diagram of an omnidirectional polar pattern. The method taught herein can be used to translate a microphone with one specific polar pattern to a microphone with a distinct polar pattern (i.e., unidirectional microphone to a bidirectional microphone, etc.). The invention includes a neural net trained on microphones with different polar pattern types, and the real time audio processing allowing the translation of a signal recorded with one polar pattern to a signal which sounds like it was recorded with a microphone having a different polar pattern.

[5400] In another embodiment, this system can be used for microphone position modeling. Condition the model by microphone distance from source, as per common recording use cases. The invention includes a neural net trained on different microphone positions, and the real time audio processing allowing signals recorded at on distance to a signal that sounds like it was recorded at a different distance.

[5500] In yet another embodiment, this system can be used for diaphragm frequency modelling. For instance, if a small diaphragm microphone is used, it may only be suited for high frequency applications. This same microphone may not be useful or ideal for low frequency applications. As such, the model can be trained by a large diaphragm microphone used for low frequency applications. The neural network can then be used to learn the difference in audio signals between the small diaphragm audio signals and the large diaphragm audio signals. The learned differences can then be applied in the future to audio signals obtained from small diaphragm microphones and convert them to audio that sounds like it was recorded using a large diaphragm microphone.

[5600] FIGs. 9A and B show a schematic diagram of the model training completed by the Digital Audio Processing System 100. The digital audio processing system 100 comprises at least a two or more microphones 103, 104, an audio analog to digital converter 112, digital audio workstation 113, and the recorded audio signals 105, 106. FIG. 9B shows the continuation of the method of the Digital Audio Processing System 100. The audio dataset 105 processed by the first microphone 103 and the audio dataset 106 processed by the second microphone 104, once recorded through an analog to digital converter 112 and into a digital audio workstation 113, can be used as digital audio inputs to the deep neural network 107. The neural network 107 can use distinct digital audio inputs 105 and 106 and learn the differences between the wave forms. The learned differences 114 between the waveforms are then saved into a database so they can be accessed later to apply the learned differences 114 to new audio from new microphones.

[5700] FIG. 10 depicts a schematic diagram of the deep neural network algorithm, as shown in FIG. 5. A microphone use-case model, or representative model is to be determined 501. This can be done via a computational processor or as a user input. The representative models include (but are not limited to): ameliorating microphone quality, polar pattern translation, microphone position modelling, singing, conversational audio, lyrical, instrumental audio, etc. Audio data from the use case is then obtained 502. At least a first and second microphone 103, 104 are then obtained 503. These microphones can be differing in some way such as quality, polar pattern, position, etc.). Any number of microphones can be used to train the data. Subsequently, a form of audio is played on a speaker and recorded on microphones 103, 104. The form audio is ideally related to the use case (i.e. conversation, music, lyrical, noise cancellation, instrumental, etc.). It is also ideal if the diaphragms of the microphones 103, 104 are aligned. The audio signals 105 and 106 can be obtained by simultaneously recording the audio on microphones 103 and 104 (see 504 of FIG. 10). As a result, at least two datasets 105, 106 are obtained from the recording captured by microphones 103, 104 (see 505 and 506 of FIG. 10). The audio dataset 505 and 506 can then be used as an input for a deep neural network 107 (see 507 of FIG. 10).

[5800] Step 507 is explained as follows. The model is divided into three parts: adaptive front-end, synthesis back-end and latent-space DNN. The architecture is designed to model nonlinear audio effects with short-term memory and is based on a parallel combination of cascade input filters, trainable wave-shaping nonlinearities, and output filters.

[5900] All convolutions are along the time dimension and all strides are of unit value. This means, during convolution, we move the filters one sample at a time. In addition, padding is done on each side of the input feature maps so that the output maintains the resolution of the input. Dilation is not introduced.

[6000] One example of the training model is depicted in Table 1.1 (shown in FIG. 12A). We can use any input frame size (for instance, the ideal frame size is between 32 and 8192, with values such as 32, 64, 256, 512,1024, 2048, 4096, 8192). This represents the number of samples in the frame of audio. The audio can be sampled with acceptable hop size ranging between 2 and 8192, with an ideally the hop size can be 256, 512, 1024. The model training sampling rate (samples per second) can range between 8-192 KHz. The new input audio may also have a sampling rate that matches that of the sampling rate used for model training.

[6100] In can be appreciated that a larger frame size will result in more frequency resolution, but less time resolution. On the other hand, a lower frame size will result in a lower frequency resolution, but a high time resolution. Different applications require varying levels of frequency resolution/time resolution. For instance, if the audio processing needed to be completed in real-time, a smaller frame size should be used. If the audio processing needed to be completed high frequency resolution, a larger frame size should be used. As such, the frame size can be moderated based on the application that needs to be achieved. In another embodiment, the DNN can be pre-set with all the ideal parameters for one application, such as for OEM (original equipment manufacturer) MEMS microphones. In another embodiment, the parameters can be left open to be chosen and set by the user.

[6200] The Adaptive front-end can comprise a convolutional encoder. It can contain two convolutional layers, one pooling layer and one residual connection. The front-end can be considered adaptive since its convolutional layers learn a filter bank for each modeling task and directly from the first microphone input audio dataset 105.

[6300] The first convolutional layer is followed by the *absolute value* as the nonlinear activation function and the second convolutional layer are locally connected (LC). This means we follow a filter bank architecture since each filter is only applied to its corresponding row in the input feature map. The later layer is followed by the *softplus* nonlinearity. The *max-pooling* layer is a moving window layer, where the maximum value within each window corresponds to the output and the positions of the maximum values are stored and used by the back-end. The operation performed by the first layer is shown in FIG. 12B (equations 1.2 and 1.3).

[6400] In equation 1.2 and 1.3, **W**1 represents the kernel matrix from the first layer, and **X**1 represents the feature map after the input audio x is convolved with **W**1. The weights **W**1 may comprise any number of one-dimensional filters having a size between (2-512), ideally 64. The residual connection **R** is equal to **X**1, which corresponds to the frequency band decomposition of the input x. This is due the output of each filter of *Conv1D* can be seen as a frequency band. The operation performed by the second layer is described by the equation 1.4 shown in FIG. 12B. Equation 1.4 (see FIG. 12B) shows an example where the filter size 128.

[6500] In equation 1.4, **X2**(^{i}) and **W2**(^{i}) are the *ith* row of the feature map **X**2 and kernel matrix **W**2, respectively. Thus, **X**2 is obtained after the LC convolution with **W**2, the weight matrix of *Conv*1*D-local*, which, in this example, has 128 filters of size 128. f2() is the *softplus* function.

[6600] The adaptive front-end performs time-domain convolutions with the first microphone 103 input audio dataset 105 and is designed to learn a latent representation for each audio effect modeling task, such . It also generates a residual connection which is used by the back-end to facilitate the synthesis of the waveform based on the specific audio effect transformation.

[6700] This differs from traditional encoding practices, where the complete input data is encoded into a latent-space, which causes each layer in the decoder to solely generate the complete desired output.

[6800] By using the *absolute value* as the activation function of the first layer and by having larger filters **W**2, we expect the front-end to learn smoother representations of the incoming audio.

[6900] Optionally, the latent-space DNN contains two fully-connected (FC) layers. Following the filter bank architecture, the first layer is based on LC layers and the second layer comprises a FC layer. The DNN modifies the latent representation **Z** into a new latent representation **Z**^ which is fed into the synthesis back-end. The first layer applies a different FC layer to each row of the matrix **Z** and the second layer is applied to each row of the output matrix from the first layer. In both layers, the number of hidden units are calculated using half the filter size, (for example, if the filter size is 128, the number of hidden units would be 64) , are followed by an activation function such as: *softplus, tanh, reLU, etc.* which can be applied to the complete latent representation rather than to the channel dimension.

[7000] The operation performed by the latent-space DNN is shown by equations 1.5 and 1.6 shown in FIG. 12B. In equations 1.5 and 1.6, **Zh**^(^{i}) is the *ith* row of the output feature map **Zh**^ of the LC layers. Likewise, **V1**(^{i}) is the ith FC layer corresponding to the weight matrix **V1** of the LC layer. **V**2 corresponds to the weights of the FC layer. The output of the max pooling operation **Z** corresponds to an optimal latent representation of the input audio.

[7100] The synthesis back-end accomplishes the nonlinear task by the following steps. First, **X2**^, the discrete approximation of **X2**, is obtained via unpooling the modified envelopes **Z**^. Then the feature map **X1**^ is the result of the element-wise multiplication of the residual connection **R** and **X2**^. This can be seen as an input filtering operation, since a different envelope gain is applied to each of the frequency band decompositions obtained in the front-end.

[7200] The deep neural network smooth adaptive activation functions (DNN-SAAF) step applies various wave-shaping nonlinearities to **X1^**. This is achieved with a processing block containing dense layers and smooth adaptive activation functions. In one embodiment, the DNN-SAAF comprises 4 fully connected layers. However, it can be appreciated that the DNN-SAAF can be any number of layers.

[7300] All fully connected layers are followed by an activation function such as tanh, *softplus, reLU, etc.* with the exception of the last layer. Locally connected SAAFs are used as the nonlinearity for the last layer. Overall, each function can be locally connected and composed of intervals ranging between 2-100 (ideally having a value of 9-25) between −1 to +1.

[7400] Finally, the deconvolution layer corresponds to the deconvolution operation, which can be implemented by transposing the first layer transform. This layer is not trainable since its kernels are transposed versions of **W**1. In this way, the back-end reconstructs the audio waveform in the same manner that the front-end decomposed it. In one embodiment, the complete waveform can be synthesized using windowing and constant overlap-add gain.

[7500] The DNN can optimize the loss function to determine the difference between the audio waveform 105 that was processed by the DNN 507 and the second microphone audio dataset 106. The first microphone audio dataset 105 is process by the neural network. The second microphone audio dataset 106, is the ideal audio dataset, and thus, is not processed by the DNN. The number of iterations can be arbitrary and will stop once the loss function is minimized.

[7600] The neural network 507 then outputs a set of learned weights for the deep neural network layers (508). Neural network architecture is implemented to learn processing (509). The learned weights **W**, or learned differences, can then be applied to new audio that has not been seen by the DNN (510). The new audio may be in the form of a pre-recorded digital signal or as a real-time audio converted to a digital signal. As a result, the new audio, which may have been recorded on a microphone of lesser quality, or microphone of non-ideal characteristics, may be upgraded to sound like audio recorded on a high quality microphone, or a microphone of ideal characteristics.

[7700] The process taught by FIG. 11 can be used to model the difference of, and upgrade speakers. In this embodiment, a reference microphone can be used to record the same audio dataset through two different loudspeakers. It can be appreciated that it is ideal to keep all other acoustic variables the same. The first speaker audio can be captured with the reference microphone, and its audio signals processed to a first dataset. The second speaker audio can simultaneously be captured with the reference microphone, and its audio signals processed to a second dataset. The neural network can be used to learn the difference between these datasets. The learned differences can then be applied to new speaker sounds in real time or recordings. It can be appreciated that the deep neural network learns transformations, and as such, it is possible to modify old recordings through conventional signal processing.

[7800] FIG. 11 depicts a schematic diagram of the digital audio processing system algorithm for upgrading a speaker. A speaker use-case model, or representative model is to be determined 601. This can be done via a computational processor or as a user input. The representative models include (but are not limited to): ameliorating speaker quality, interference pattern, speaker position modelling, singing, conversational audio, lyrical, instrumental audio, etc. Audio data from the use case is then obtained 602. At least a first and second speakers are then obtained 603. These speakers can be differing in some way such as quality, interference pattern, position, etc.). Any number of speakers can be used to train the data. Subsequently, a form of audio is played on a microphone and recorded on at least two speakers. The form audio is ideally related to the use case (i.e. conversation, music, lyrical, noise cancellation, instrumental, etc.). The audio signals can be obtained by simultaneously recording the audio on the two speakers (see 604). As a result, at least two datasets are obtained from the recording captured by the speakers (see 605 and 606). The differing audio datasets can then be used as an input for a deep neural network 107 (see 607). The neural network outputs a set of learned weights for the deep neural network layers (608). Neural network architecture is implemented to learn processing (609). The learned weights, or learned differences, can then be applied to new audio that has not been seen by the DNN (610). The new audio may be in the form of a pre-recorded digital signal or on real-time audio converted to a digital signal. As a result, the new audio, which may be playing on a lesser quality speaker, or speaker of non-ideal characteristics, may be upgraded to sound like audio being played on a high quality speaker, or a speaker of ideal characteristics.

[7900] Although the above description includes reference to certain specific embodiments, various modifications thereof will be apparent to those skilled in the art. Any examples provided herein are included solely for the purpose of illustration and are not intended to be limiting in any way. Any drawings provided herein are solely for the purpose of illustrating various aspects of the description and are not intended to be drawn to scale or to be limiting in any way. The scope of the claims appended hereto should not be limited by the preferred embodiments set forth in the above description but should be given the broadest interpretation consistent with the present specification as a whole. The disclosures of all prior art recited herein are incorporated herein by reference in their entirety.

WE CLAIM:

1. A method of improving an audio signal comprising:

outputting an audio waveform from a sound source;

capturing the audio waveform from a first microphone and capturing the audio waveform from a second microphone capsule aligned beside the first microphone; and

sending the captured audio waveforms to a digital audio processing system having a neural network;

such that the neural network is configured to learn differences between the first audio waveform and the second audio waveform; and

such that the audio signals processed from the first microphone is differing from the audio signal processed from second microphone.

2. The method of claim 1, wherein the method further comprises: applying the learned differences to a third audio waveform recorded from a third microphone; such that the third microphone has similar characteristics to the first microphone.

3. The method of claim 1 wherein the sound source comprises curated data set consisting of test signals, representative audio signals.

4. The method of claim 1 wherein the first microphone has a non-ideal set of characteristics, and the second microphone has an ideal set of characteristics for a specific function.

5. The method of claim 4, wherein the specific function is selected from at least one of the following: conversation, lyrical, music, noise cancellation, and instrumental function.

6. The method of claim 2, wherein the third microphone is located on a mobile device.

7. The method of claim 6, wherein the third microphone records audio from a telephone conversation such that the digital audio processing system processes the conversation in real-time.

8. The method of claim 7, wherein the digital audio processing system is located on an application on the mobile device.

9. The method of claim 1, wherein the method is used for polar pattern translation, wherein the first microphone comprises a first polar pattern from one of a unidirectional, bidirectional, and omnidirectional pattern, and the second microphone comprises a second polar pattern, from one of a unidirectional, bidirectional, and omnidirectional microphone.

10. The method of claim 9, wherein the third microphone comprises the first polar pattern and wherein the learned differences between the first and second microphone are applied to a third audio waveform recorded from the third microphone to match the second polar pattern.

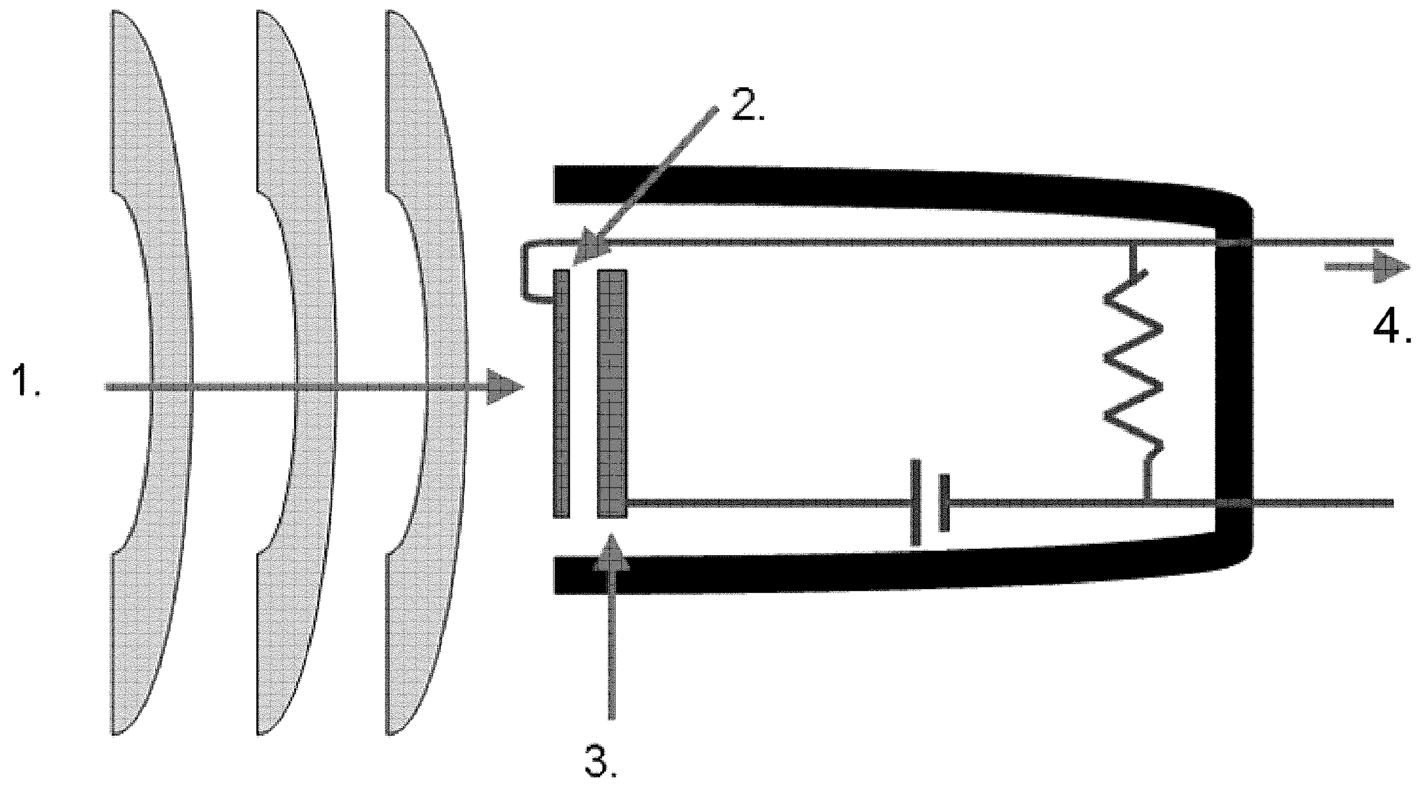


FIG. 1
PRIOR ART

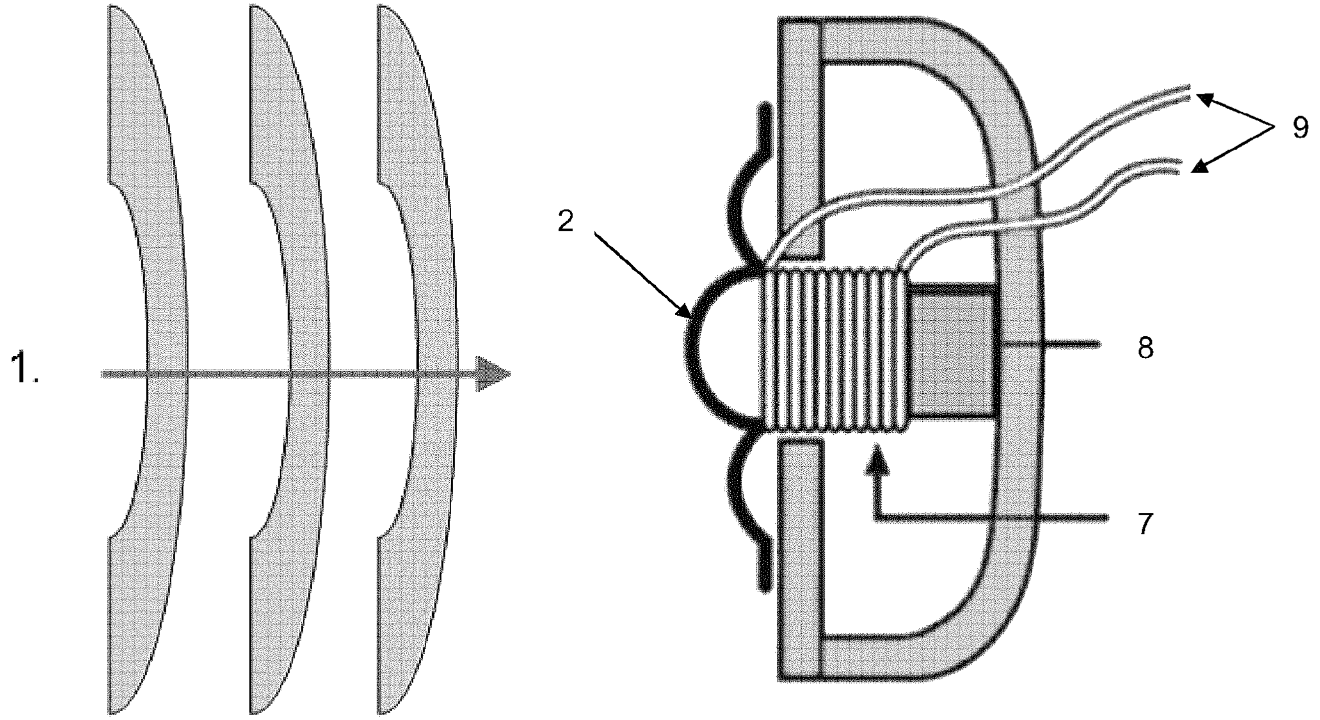


FIG. 2
PRIOR ART

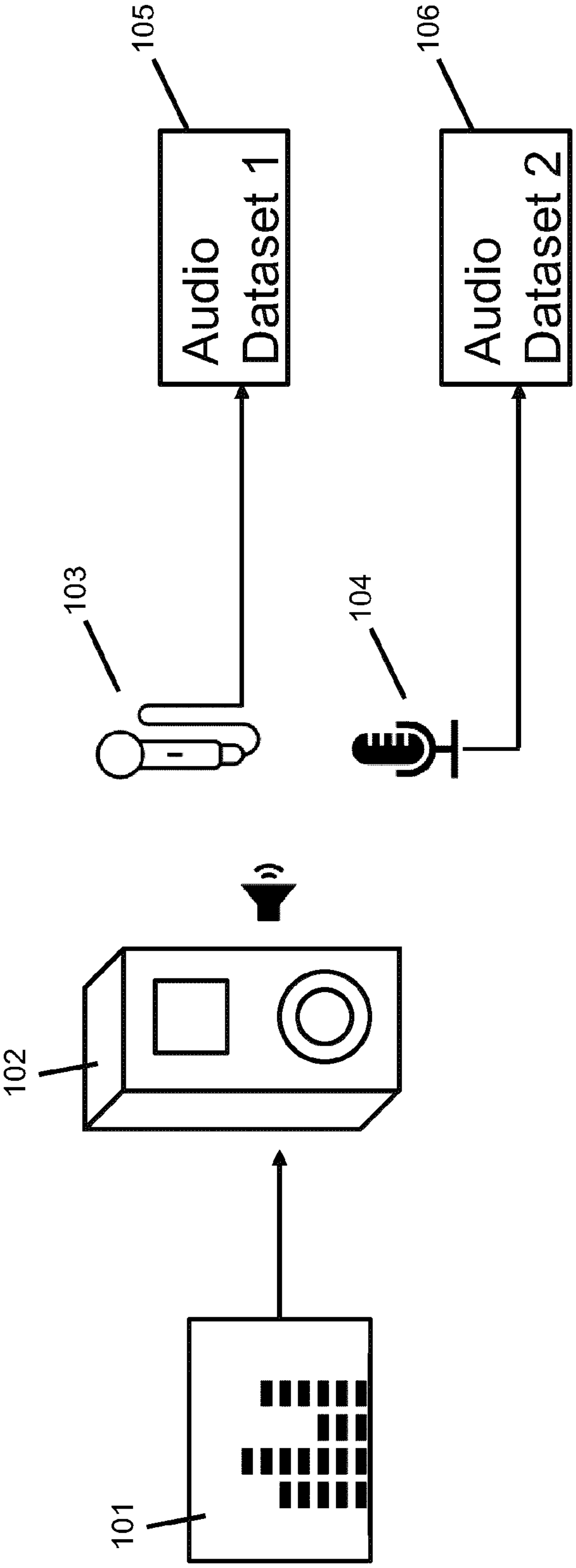


FIG. 3

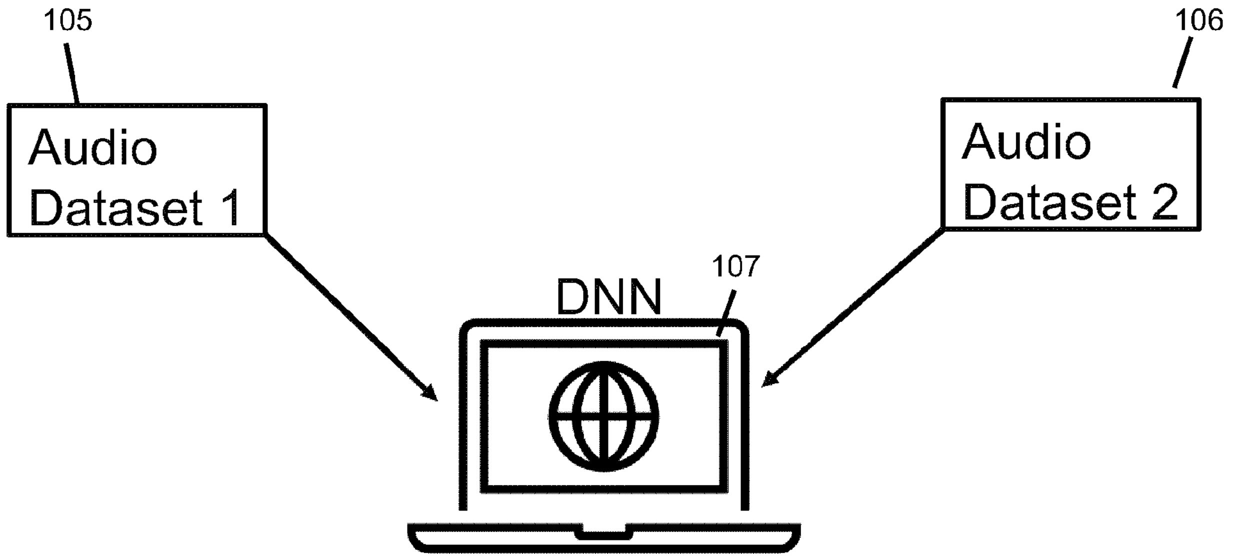
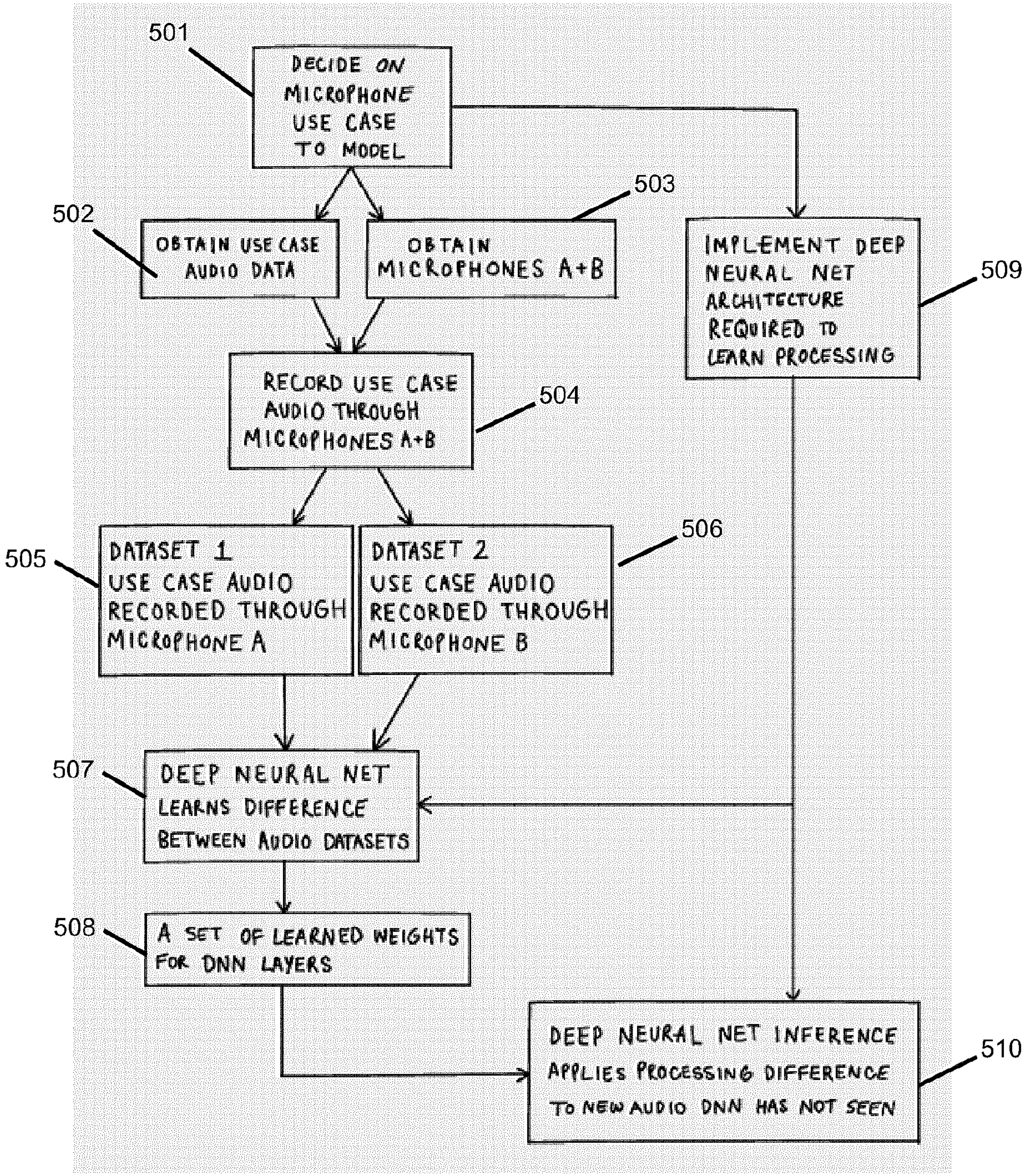


FIG. 4

FIG. 5


501
DECIDE ON MICROPHONE USE CASE TO MODEL
502
OBTAIN USE CASE AUDIO DATA
503
OBTAIN MICROPHONES A+B
504
RECORD USE CASE AUDIO THROUGH MICROPHONES A+B
505
DATASET 1 USE CASE AUDIO RECORDED THROUGH MICROPHONE A
506
DATASET 2 USE CASE AUDIO RECORDED THROUGH MICROPHONE B
507
DEEP NEURAL NET LEARNS DIFFERENCE BETWEEN AUDIO DATASETS
508
A SET OF LEARNED WEIGHTS FOR DNN LAYERS
509
IMPLEMENT DEEP NEURAL NET ARCHITECTURE REQUIRED TO LEARN PROCESSING
510
DEEP NEURAL NET INFERENCE APPLIES PROCESSING DIFFERENCE TO NEW AUDIO DNN HAS NOT SEEN

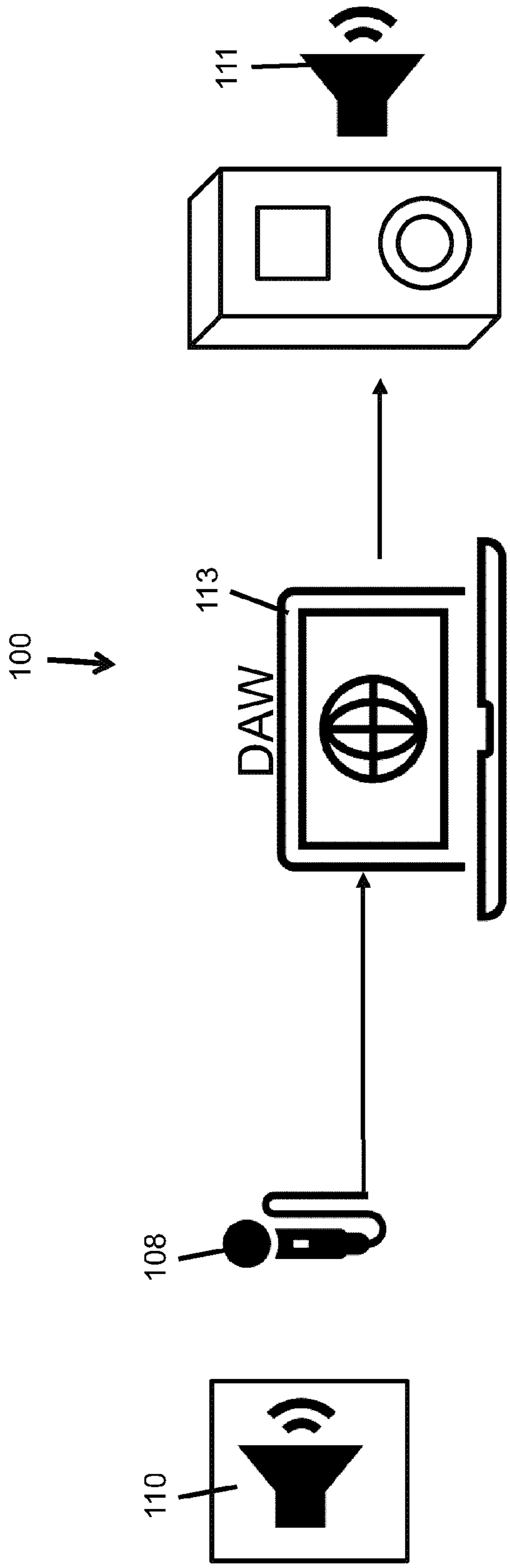


FIG. 6

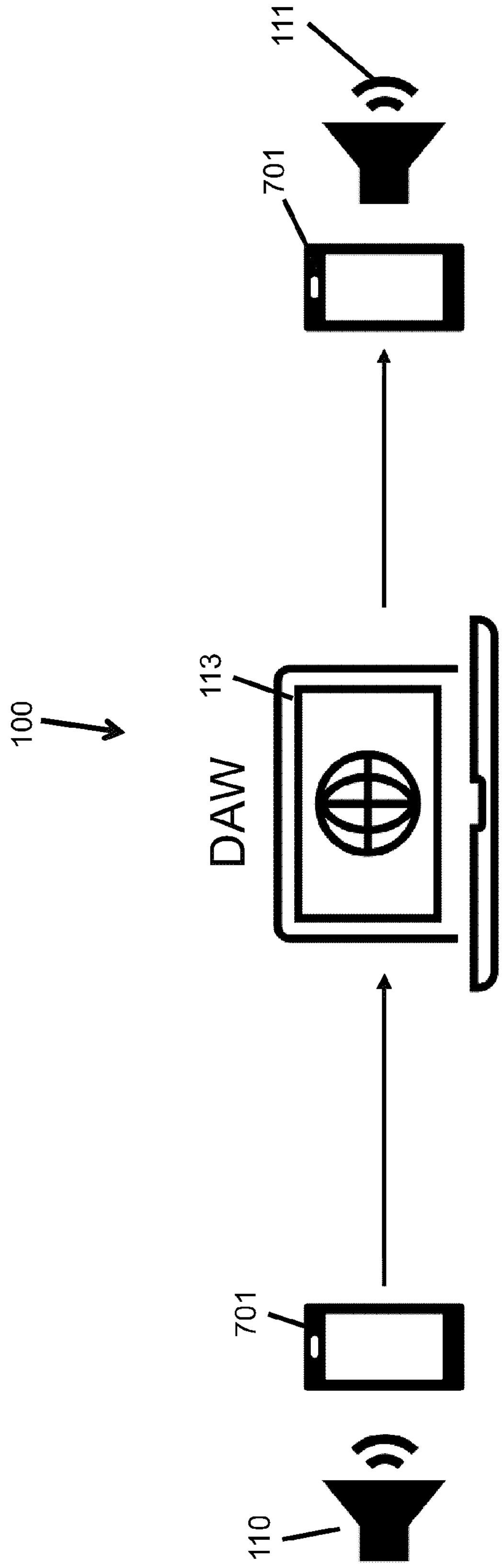


FIG. 7

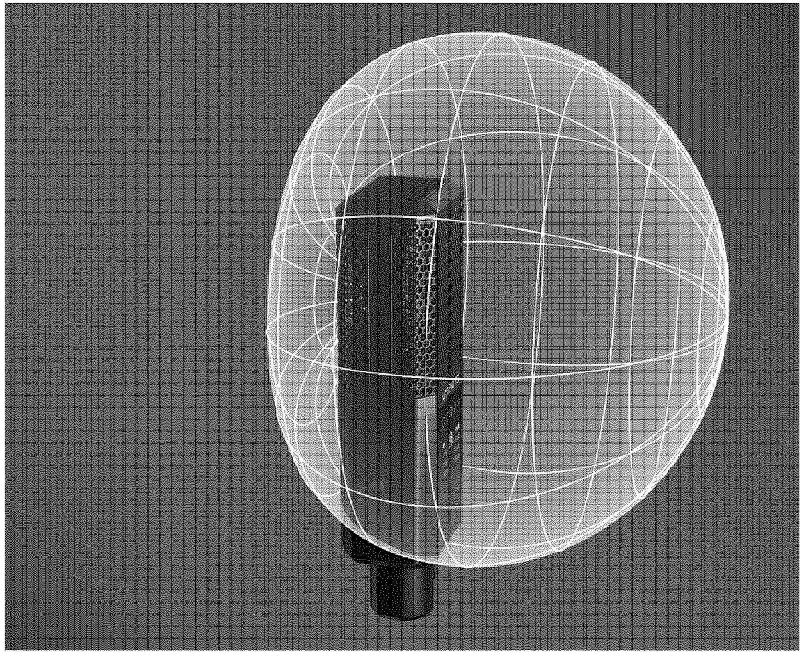


FIG. 8A (unidirectional)

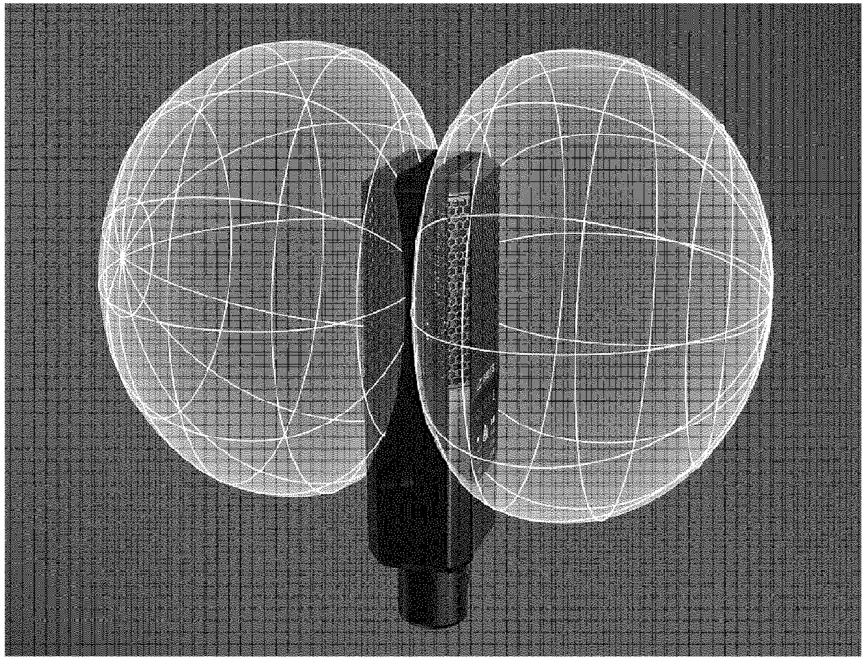


FIG. 8B (bidirectional)

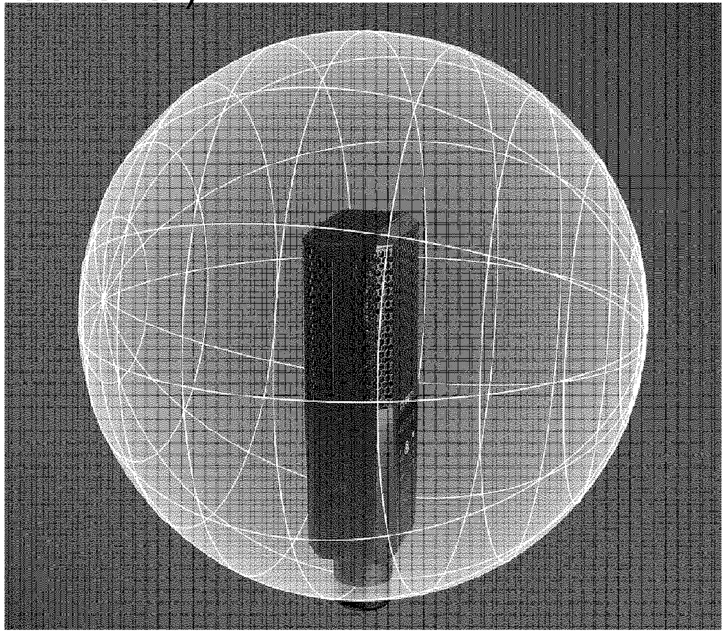


FIG. 8C (omnidirectional)

FIG. 9A

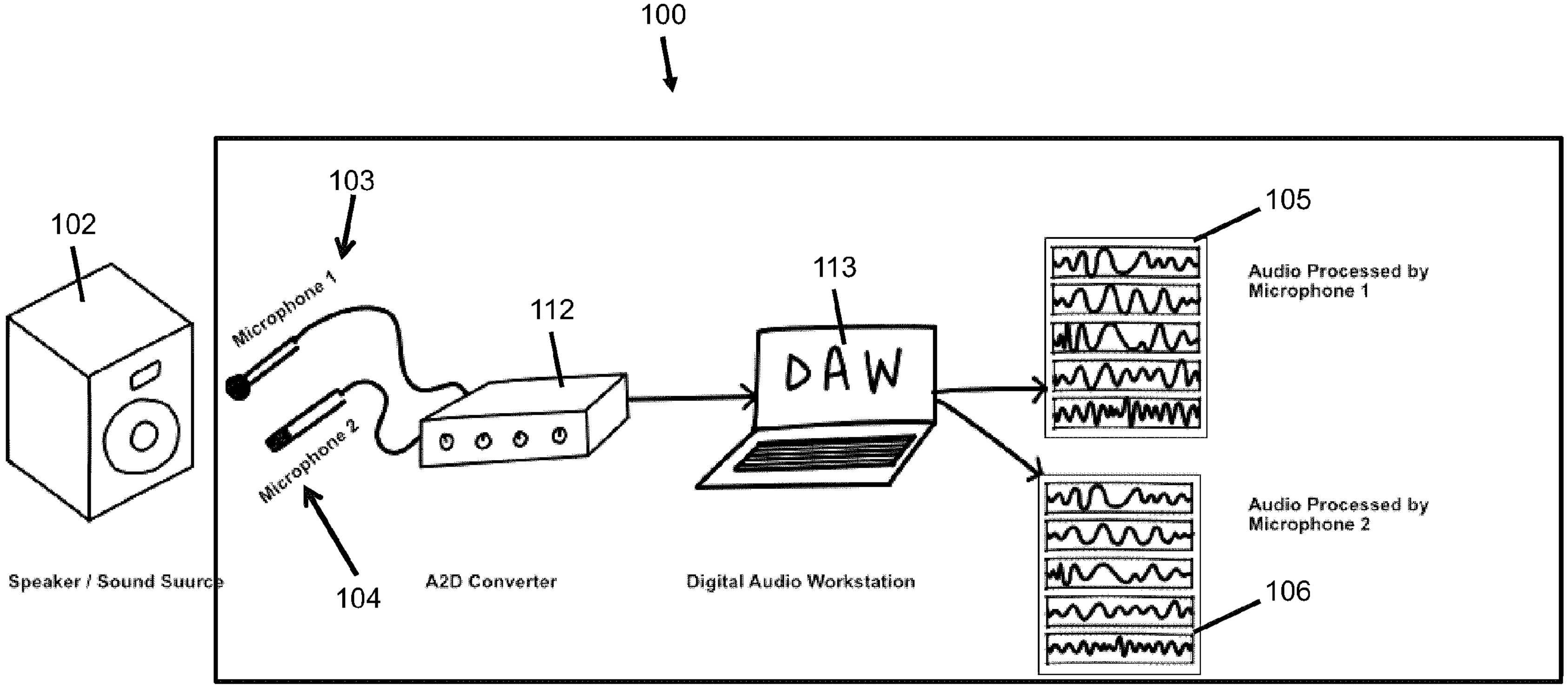

100
102
103
Microphone 1
104
Microphone 2
112
113
DAW
105
Audio Processed by Microphone 1
106
Audio Processed by Microphone 2
Speaker / Sound Suurce
A2D Converter
Digital Audio Workstation

FIG. 9B

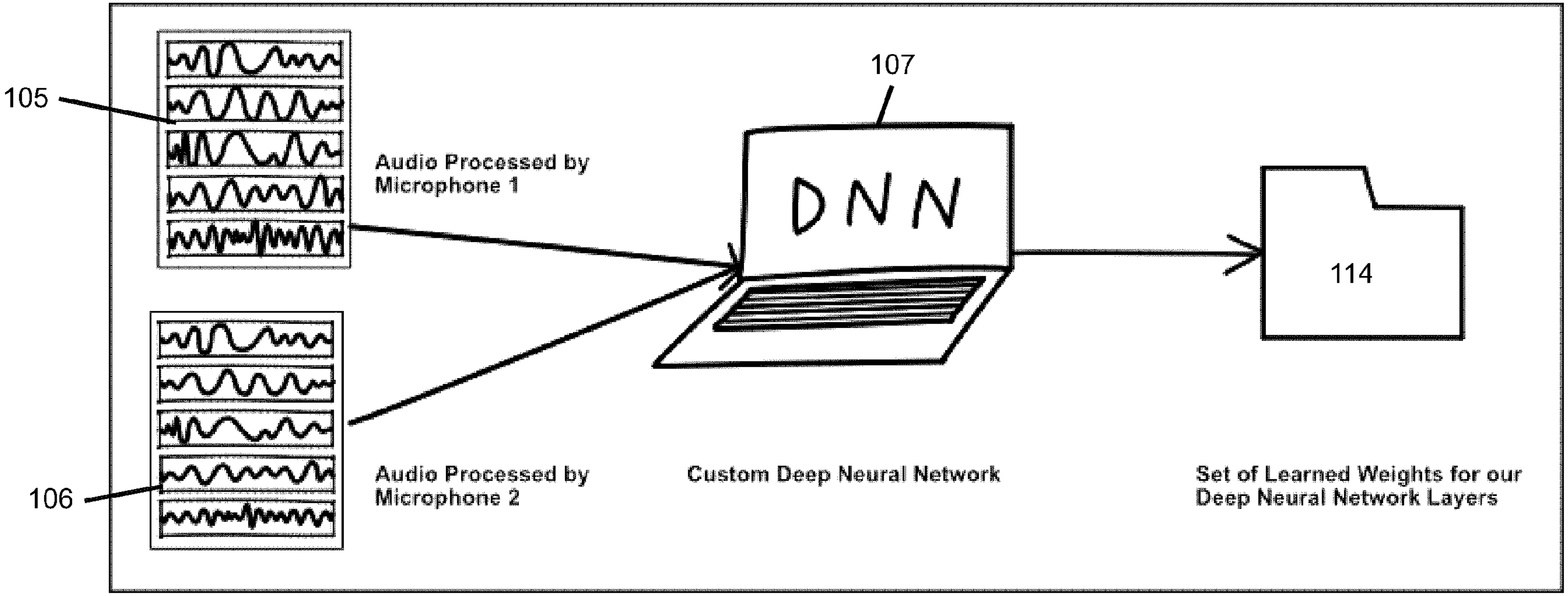
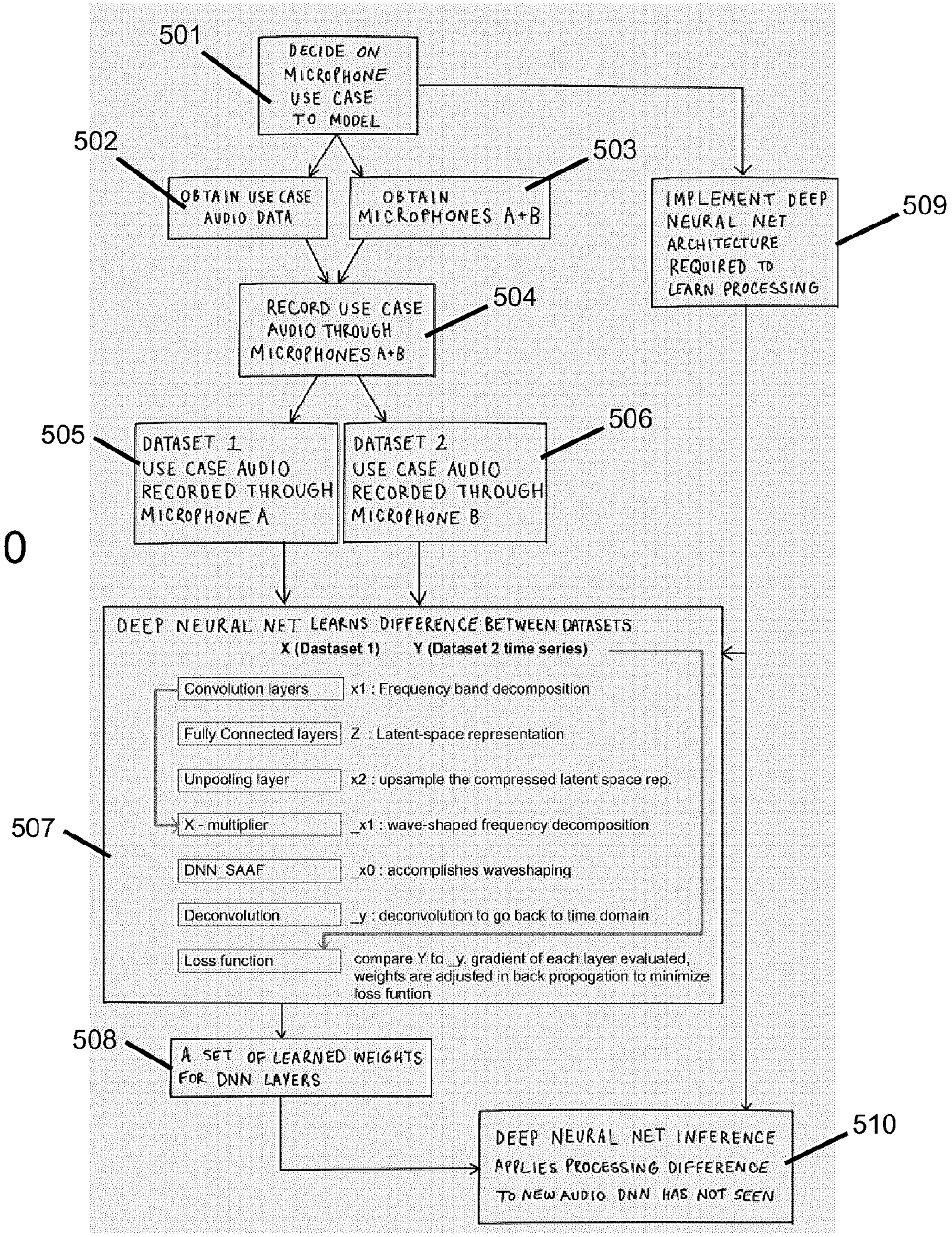

105
106
Audio Processed by Microphone 1
Audio Processed by Microphone 2
107
DNN
Custom Deep Neural Network
114
Set of Learned Weights for our Deep Neural Network Layers

FIG. 10


501
DECIDE ON MICROPHONE USE CASE TO MODEL
502
OBTAIN USE CASE AUDIO DATA
503
OBTAIN MICROPHONES A+B
509
IMPLEMENT DEEP NEURAL NET ARCHITECTURE REQUIRED TO LEARN PROCESSING
504
RECORD USE CASE AUDIO THROUGH MICROPHONES A+B
505
DATASET 1 USE CASE AUDIO RECORDED THROUGH MICROPHONE A
506
DATASET 2 USE CASE AUDIO RECORDED THROUGH MICROPHONE B
507
DEEP NEURAL NET LEARNS DIFFERENCE BETWEEN DATASETS
X (Dastaset 1)
Y (Dataset 2 time series)
Convolution layers
x1 : Frequency band decomposition
Fully Connected layers
Z : Latent-space representation
Unpooling layer
x2 : upsample the compressed latent space rep.
X - multiplier
_x1 : wave-shaped frequency decomposition
DNN_SAAF
_x0 : accomplishes waveshaping
Deconvolution
_y : deconvolution to go back to time domain
Loss function
compare Y to _y. gradient of each layer evaluated, weights are adjusted in back propogation to minimize loss funtion
508
A SET OF LEARNED WEIGHTS FOR DNN LAYERS
510
DEEP NEURAL NET INFERENCE APPLIES PROCESSING DIFFERENCE TO NEW AUDIO DNN HAS NOT SEEN

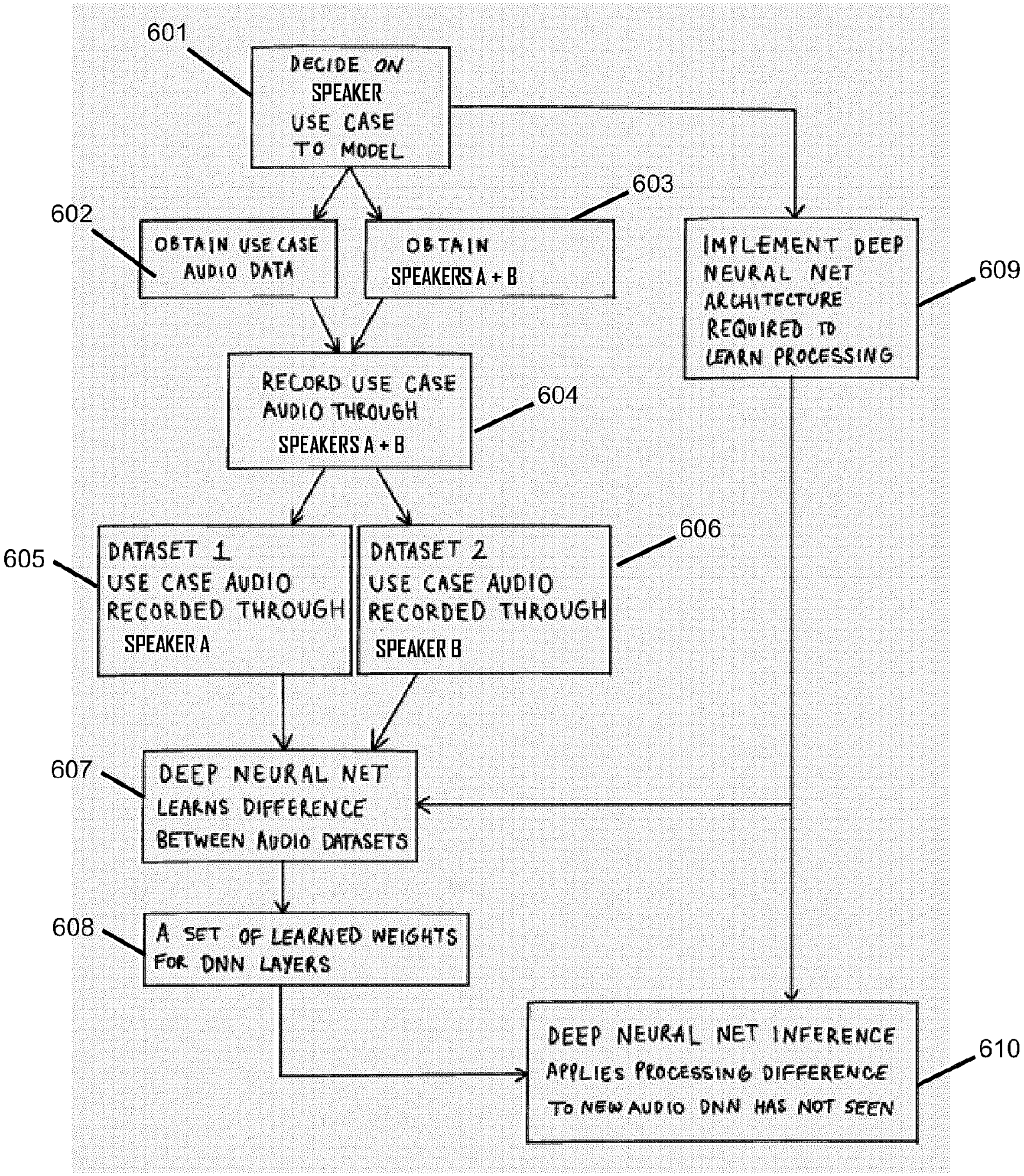

601
DECIDE ON SPEAKER USE CASE TO MODEL
602
OBTAIN USE CASE AUDIO DATA
603
OBTAIN SPEAKERS A + B
604
RECORD USE CASE AUDIO THROUGH SPEAKERS A + B
605
DATASET 1 USE CASE AUDIO RECORDED THROUGH SPEAKER A
606
DATASET 2 USE CASE AUDIO RECORDED THROUGH SPEAKER B
607
DEEP NEURAL NET LEARNS DIFFERENCE BETWEEN AUDIO DATASETS
608
A SET OF LEARNED WEIGHTS FOR DNN LAYERS
609
IMPLEMENT DEEP NEURAL NET ARCHITECTURE REQUIRED TO LEARN PROCESSING
610
DEEP NEURAL NET INFERENCE APPLIES PROCESSING DIFFERENCE TO NEW AUDIO DNN HAS NOT SEEN

FIG. 11

Table 1.1: Detailed architecture of *CAFx* with an input frame size of 1024 samples.

Layer	Output shape	Weights	Output
Input	(1024, 1)	.	x
Conv1D	(1024, 128)	128(64)	X_1
Residual	(1024, 128)	.	R
Abs	(1024, 128)	.	.
Conv1D-Local	(1024, 128)	128(128)	X_2
MaxPooling	(64, 128)	.	Z
Dense-Local	(128, 64)	64(128)	.
Dense	(128, 64)	64	$\hat{Z}$
Unpooling	(1024, 128)	.	$\hat{X}_2$
$R \times \hat{X}_2$	(1024, 128)	.	$\hat{X}_1$
Dense	(1024, 128)	128	.
Dense	(1024, 64)	64	.
Dense	(1024, 64)	64	.
Dense	(1024, 128)	128	.
SAAF	(1024, 128)	128(25)	$\hat{X}_0$
deConv1D	(1024, 1)	.	$\hat{y}$

FIG. 12A

$$X_1 = x * W_1 \qquad (1.2)$$

$$R = X_1 \qquad (1.3)$$

$$X_2^{(i)} = f_2(|X_1^{(i)}| * W_2^{(i)}),\ \forall i \in [1, 128] \qquad (1.4)$$

$$\hat{Z}_h^{(i)} = f_h(Z^{(i)} \cdot V_1^{(i)}),\ \forall i \in [1, 64] \qquad (1.5)$$

$$\hat{Z} = f_h(\hat{Z}_h \cdot V_2) \qquad (1.6)$$

FIG. 12B

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CA2022/051637

A. CLASSIFICATION OF SUBJECT MATTER

IPC: ***H04R 3/00*** (2006.01) , ***G10L 25/30*** (2013.01)

CPC: **H04R 3/00** (2020.01) , G10L 25/30 (2020.01)

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC: ***H04R 3/00*** (2006.01) , ***G10L 25/30*** (2013.01)
CPC: **H04R 3/00** (2020.01) , G10L 25/30 (2020.01)

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

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

Databases: CPD, Questel Orbit, Google Patents
Keywords: audio signal improvement, audio waveform, capture, microphone, audio processing system, neural network, data, mobile device, record, application, polar pattern

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US2020211540(A1) 2 July 2020 (02-07-2020) by MacConnell et al. ** see abstract, entire application**	1-10
A	US2020367810(A1) 26 Nov. 2020(26-11-2020) by Shouldice et al. ** see abstract, entire application **	1-10
A	US2019261914(A1) 29 Aug. 2019 (29-08-2019) by Davis et al. ** see abstract, entire application**	1-10

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

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"D"	document cited by the applicant in the international application	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"E"	earlier application or patent but published on or after the international filing date	"&"	document member of the same patent family
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)		
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 11 February 2023 (11-02-2023)	Date of mailing of the international search report 15 February 2023 (15-02-2023)
Name and mailing address of the ISA/CA Canadian Intellectual Property Office Place du Portage I, C114 - 1st Floor, Box PCT 50 Victoria Street Gatineau, Quebec K1A 0C9 Facsimile No.: 819-953-2476	Authorized officer Karen Oprea (819) 639-8255

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CA2022/051637

Patent Document Cited in Search Report	Publication Date	Patent Family Member(s)	Publication Date
US2020211540A1	02 July 2020 (02-07-2020)	US2020211540A1	02 July 2020 (02-07-2020)
		CN113228162A	06 August 2021 (06-08-2021)
		EP3903305A1	03 November 2021 (03-11-2021)
		WO2020139724A1	02 July 2020 (02-07-2020)
US2020367810A1	26 November 2020 (26-11-2020)	US2020367810A1	26 November 2020 (26-11-2020)
		CN111655125A	11 September 2020 (11-09-2020)
		EP3727134A1	28 October 2020 (28-10-2020)
		EP3727134B1	25 January 2023 (25-01-2023)
		JP2021508279A	04 March 2021 (04-03-2021)
		JP7202385B2	11 January 2023 (11-01-2023)
		KR20200104341A	03 September 2020 (03-09-2020)
		WO2019122412A1	27 June 2019 (27-06-2019)
US2019261914A1	29 August 2019 (29-08-2019)	US2019261914A1	29 August 2019 (29-08-2019)
		CN105393252A	09 March 2016 (09-03-2016)
		CN105393252B	19 April 2019 (19-04-2019)
		EP2987106A1	24 February 2016 (24-02-2016)
		EP2987106A4	14 December 2016 (14-12-2016)
		US2015005644A1	01 January 2015 (01-01-2015)
		US9414780B2	16 August 2016 (16-08-2016)
		US2015006186A1	01 January 2015 (01-01-2015)
		US9445763B2	20 September 2016 (20-09-2016)
		US2015003699A1	01 January 2015 (01-01-2015)
		US9504420B2	29 November 2016 (29-11-2016)
		US2017143249A1	25 May 2017 (25-05-2017)
		US10219736B2	05 March 2019 (05-03-2019)
		US2014313303A1	23 October 2014 (23-10-2014)
		US2014316235A1	23 October 2014 (23-10-2014)
		US2014378810A1	25 December 2014 (25-12-2014)
		US2015003698A1	01 January 2015 (01-01-2015)
		US2015005640A1	01 January 2015 (01-01-2015)
		WO2014172671A1	23 October 2014 (23-10-2014)