



(51) International Patent Classification:

H04N 19/46 (2014.01) *H04N 21/8547* (2011.01)
H04N 19/68 (2014.01) *G06N 3/045* (2023.01)
H04N 19/85 (2014.01) *G06N 3/0464* (2023.01)

(21) International Application Number:

PCT/US2024/054425

(22) International Filing Date:

04 November 2024 (04.11.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/596,485 06 November 2023 (06.11.2023) US
 23212036.0 24 November 2023 (24.11.2023) EP

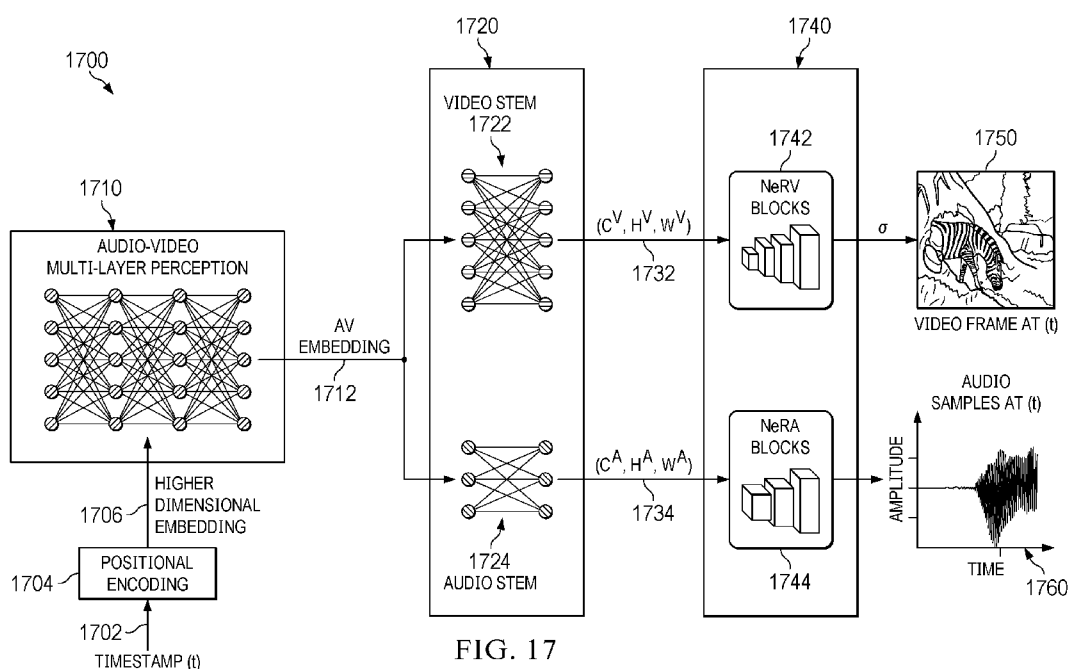
(71) Applicant: **DOLBY LABORATORIES LICENSING CORPORATION** [US/US]; 1275 Market Street, San Francisco, California 94103 (US).

(72) Inventors: **SINGH, Praneet**; 475 South St, Apt 306, Lafayette, Indiana 47901 (US). **CHOUDHURY, Anustup Kumar Atanu**; c/o Dolby Laboratories, Inc., 1275 Market Street, San Francisco, California 94103 (US). **SU, Guan-Ming**; c/o Dolby Laboratories, Inc., 1275 Market Street, San Francisco, California 94103 (US).

(74) Agent: **KONSTANTINIDES, Konstantinos** et al.; Dolby Laboratories, Inc., Intellectual Property Group, 1275 Market Street, San Francisco, California 94103 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,

(54) Title: MULTIMEDIA REPRESENTATION USING NEURAL FIELD



(57) **Abstract:** Methods and apparatus for representing multimedia signals using neural field networks. According to an example embodiment, a method of jointly representing an audio signal and a video signal using a neural field includes applying positional encoding to a time stamp corresponding to a video frame of the video signal and further corresponding to an associated segment of the audio signal to generate a respective higher-dimensional embedding. The method further includes reconstructing the video frame of the video signal and generating a corresponding set of audio samples representing the associated segment of the audio signal by applying the higher-dimensional embedding to a neural-field network trained to represent a sequence of video frames of the video signal and an associated plurality of segments of the audio signal.



HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

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

Published:

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

MULTIMEDIA REPRESENTATION USING NEURAL FIELD

1. Cross-Reference to Related Applications

[0001] This application claims the benefit of priority from U.S. Provisional Application Ser. No. 63/596,485, filed on 6 November 2023, and European Patent Application No. EP23212036.0, filed on 24 November 2023, each of which is incorporated by reference herein in its entirety.

2. Field of the Disclosure

[0002] Various example embodiments relate to neural field representations of multimedia.

3. Background

[0003] Advances in machine learning have led to the use of methods employing coordinate-based neural networks for solving certain visual-computing problems. Such neural networks, often referred to as “neural fields,” parameterize physical properties of scenes and objects across space and time. Example applications of neural fields include 3D shape and image synthesis, animation of human bodies, and pose estimation. Additional applications of neural fields are currently being actively developed.

BRIEF SUMMARY OF SOME SPECIFIC EMBODIMENTS

[0004] Disclosed herein are various embodiments of methods and apparatus for representing audio and video signals using neural field networks. Some embodiments are directed to using neural fields to represent Short-Time Fourier Transform (STFT) Magnitude and Phase spectrograms of audio signals. Some other embodiments are directed to time-stamp-based neural field representations of audio signals. Further embodiments are directed to jointly representing an audio signal and a video signal using a neural field network. In some of the latter embodiments, we use a time-stamp-based approach and employ a modality-unifying audio-video Multi-Layer Perceptron (MLP) that brings together audio and video information in a unified framework. Modality-dependent reconstruction branches coupled to this MLP are then used to independently reconstruct audio and video signals using embeddings generated by the MLP. Various embodiments may also employ swish and sine activation functions and higher dimensional positional encodings.

[0005] Findings disclosed herein indicate that joint audio-video representations tend to outperform individual modality representations in terms of reconstruction performance. In addition, some of the joint audio-video representations demonstrate versatility beyond integrating audio and video data, for example, by being upgradable for incorporating additional modalities, such as textual data and/or sensor inputs. Some of the disclosed neural field architectures beneficially exploit inter-modal redundancies, thereby making such architectures well-suited for multi-modal signal-compression applications.

[0006] According to an example embodiment, provided is a method of representing an audio signal using a neural field, the method comprising: transforming the audio signal into a pair of 2D spectrograms including a 2D magnitude spectrogram and a corresponding 2D phase spectrogram, a first dimension of the 2D spectrograms representing time, a second dimension of the 2D spectrograms representing frequency; and training a neural-field network to represent the pair of 2D spectrograms, the neural-field network being configured to output a magnitude value and a phase value corresponding to the audio signal in response to an input of a 2D coordinate, the training producing a set of fixed parameters for the neural field network.

[0007] According to another example embodiment, provided is a method of representing an audio signal using a neural field, the method comprising: generating a pair of 2D spectrograms including a 2D magnitude spectrogram and a corresponding 2D phase spectrogram by testing a neural-field network that has been trained to output a magnitude value and a phase value corresponding to the audio signal in response to an input of a 2D coordinate, a first dimension of the 2D spectrograms representing time, a second dimension of the 2D spectrograms representing frequency; and reconstructing the audio signal based on the pair of 2D spectrograms.

[0008] According to yet another example embodiment, provided is a method of representing an audio signal using a neural field, the method comprising: applying positional encoding to a time stamp corresponding to a segment of the audio signal to generate a respective higher-dimensional embedding; and training a neural-field network to represent a plurality of segments of the audio signal, the neural-field network being configured to output a respective set of audio samples representing a respective one of the plurality of segments of the audio signal in response to an input of the respective higher-dimensional embedding, the training producing a set of fixed parameters for the neural field network.

[0009] According to yet another example embodiment, provided is a method of representing an audio signal using a neural field, the method comprising: applying positional encoding to a time stamp corresponding to a segment of the audio signal to generate a higher-dimensional embedding; and generating a set of audio samples representing the segment of the audio signal by applying the higher-dimensional embedding to a neural-field network trained to represent a plurality of segments of the audio signal, the neural-field network being configured to output a respective set of audio samples representing a respective one of the plurality of segments of the audio signal in response to an input of a respective higher-dimensional embedding.

[0010] According to yet another example embodiment, provided is a method of jointly representing an audio signal and a video signal using a neural field, the method comprising: applying positional encoding to a time stamp corresponding to a video frame of the video signal and an associated segment of the audio signal to generate a respective higher-dimensional embedding; and training a neural-field network to represent a sequence of video frames of the video signal and an associated plurality of segments of the audio signal, the neural-field network being configured to output a respective video frame of the sequence and a respective set of audio samples representing a respective one of the associated plurality of segments in response to an input of the respective higher-dimensional embedding, the training producing a set of fixed parameters for the neural field network.

[0011] According to yet another example embodiment, provided is a method of jointly representing an audio signal and a video signal using a neural field, the method comprising: applying positional encoding to a time stamp corresponding to a video frame of the video signal and further corresponding to an associated segment of the audio signal to generate a respective higher-dimensional embedding; and reconstructing the video frame of the video signal and generating a corresponding set of audio samples representing the associated segment of the audio signal by applying the higher-dimensional embedding to a neural-field network trained to represent a sequence of video frames of the video signal and an associated plurality of segments of the audio signal, the neural-field network being configured to output a respective video frame of the sequence and a respective set of audio samples representing a respective one of the associated plurality of segments in response to an input of the respective higher-dimensional embedding, the training producing a set of fixed parameters for the neural field network.

[0012] According to yet another example embodiment, provided is a method of jointly representing a first multimedia signal and a second multimedia signal using a neural field, the method comprising: applying positional encoding to a time stamp corresponding to a portion of the second multimedia signal and an associated segment of the first multimedia signal to generate a
5 respective higher-dimensional embedding; and training a neural-field network to represent a sequence of portions of the second multimedia signal and an associated plurality of segments of the first multimedia signal, the neural-field network being configured to output a respective portion of the sequence and a respective set of samples representing a respective one of the associated plurality of segments in response to an input of the respective higher-dimensional embedding, the training
10 producing a set of fixed parameters for the neural field network.

[0013] According to yet another example embodiment, provided is a non-transitory computer-readable medium storing instructions that, when executed by an electronic processor, cause the electronic processor to perform operations comprising any of the above methods.

BRIEF DESCRIPTION OF THE DRAWINGS

15 [0014] Other aspects, features, and benefits of various disclosed embodiments will become more fully apparent, by way of example, from the following detailed description and the accompanying drawings, in which:

[0015] FIG. 1 is a block diagram illustrating building blocks that can be used to parameterize a signal using neural fields according to some embodiments.

20 [0016] FIG. 2 is a block diagram illustrating a multilayer perceptron (MLP) that can be used to implement a neural field according to an embodiment.

[0017] FIG. 3 is a block diagram illustrating a Convolutional Neural Network (CNN) that can be used to implement a neural field according to another example.

25 [0018] FIG. 4 is a block diagram illustrating operations performed in a convolutional layer of the CNN of FIG. 3 according to an example.

[0019] FIG. 5 is a flowchart illustrating a Frechet-Audio-Distance (FAD) workflow according to one embodiment.

[0020] FIG. 6 is a block diagram illustrating a Neural Field architecture used for providing Neural Field representations for videos according to one example.

[0021] FIGS. 7A-7B graphically illustrate window functions used in Short-Time Fourier Transform (STFT) according to some examples.

5 [0022] FIG. 8 is a block diagram illustrating a workflow for generating spectrograms via the STFT according to one example.

[0023] FIG. 9 is a block diagram illustrating a neural-fields-based representation of an audio signal using the corresponding STFT magnitude and phase spectrograms according to one embodiment.

10 [0024] FIGS. 10A-10C graphically illustrate audio signal reconstruction results obtained using the neural-fields-based representation of FIG. 9 according to one example.

[0025] FIG. 11 is a block diagram of a time-stamp-based representation of an audio signal according to one embodiment.

15 [0026] FIGS. 12A-12B graphically illustrate audio signal reconstruction results obtained using the time-stamp-based representation of FIG. 11 according to one example.

[0027] FIG. 13 is a block diagram of a time-stamp-based representation of an audio signal according to another embodiment.

[0028] FIG. 14 is a block diagram illustrating a Neural Representation for Audio (NeRA) block used in the time-stamp-based representation of FIG. 13 according to one example.

20 [0029] FIGS. 15A-15B graphically illustrate audio signal reconstruction results obtained using the time-stamp-based representation of FIG. 13 according to one example.

[0030] FIG. 16 is a table comparing the performance of the time-stamp-based representations of FIGS. 11 and 13 according to various examples.

25 [0031] FIG. 17 is a block diagram illustrating a joint Neural Field representation of video and audio signals according to one embodiment.

[0032] FIG. 18 is a block diagram illustrating a joint Neural Field representation of video and audio signals according to another embodiment.

[0033] FIG. 19 is a block diagram illustrating a computing device used to implement Neural Field representations according to various embodiments.

5

DETAILED DESCRIPTION

[0034] This disclosure and aspects thereof can be embodied in various forms, including hardware, devices or circuits controlled by computer-implemented methods, computer program products, computer systems and networks, user interfaces, and application programming interfaces; as well as hardware-implemented methods, signal processing circuits, memory arrays, application specific integrated circuits (ASICs), field programmable gate arrays (FPGAs), and the like. The foregoing is intended solely to give a general idea of various aspects of the present disclosure and does not limit the scope of the disclosure in any way.

[0035] In the following description, numerous details are set forth, such as optical device configurations, timings, operations, and the like, in order to provide an understanding of one or more aspects of the present disclosure. It will be readily apparent to one skilled in the art that these specific details are merely exemplary and not intended to limit the scope of this application.

[0036] Neural Field representations have many applications across various domains, such as volumetric video coding, spatial audio rendering, and audio/video compression. The attractiveness of neural fields for such applications lies in their capability to encompass complete signals, enabling coordinate-based signal reconstruction through utilization of model weights. Additionally, timestamp-based signal reconstruction can be used for video signals. In some examples, individual machine learning models are employed to represent distinct modalities. In some cases, model weights alone allow for a complete signal reconstruction based on a time-stamp index, denoted as t . However, effective multi-modal neural field representations that seamlessly integrate multiple modalities are less developed. For example, the development of Neural Field representations for images and video appear to receive more attention from the developers than Neural Field representations for audio.

[0037] Accordingly, some embodiments disclosed herein are directed at providing neural field representations for audio signals. Example features of such embodiments include:

- Short-Time Fourier Transform (STFT) Magnitude and Phase spectrogram representations using Neural Fields.
- Time-stamp based Neural Field representations in order to allow for association of audio with other modalities, such as video signals. Such representations may be based on:

- Using the Multi-Layer Perceptron (MLP);
- A combined MLP and Convolutional Neural Network (CNN) based architecture, hereafter referred to as Neural Representation for Audio (NeRA).

10 [0038] Additional embodiments disclosed herein are directed at providing multi-modal neural field representations designed to jointly represent and reconstruct both audio and video data using time-stamp coordinates. Example features of such embodiments include:

- Modality-Unifying Audio-Video Multi-Layer Perceptron (MLP): This MLP acts as a unifying framework bringing together audio and video information.
- Modality-Dependent Convolutional Reconstruction Branches: These branches utilize embeddings obtained from the MLP to independently reconstruct audio and video signals by leveraging time-stamp indices, e.g., using Neural Representation for Video (NeRV) and a Neural Representation for Audio (NeRA).

20 [0039] Further embodiments disclosed herein are directed at utilizing swish/sine activations along with higher dimensional positional encodings to improve both individual and joint representations and reconstructions of audio and video.

[0040] At least some embodiments of a joint representation disclosed herein beneficially exhibit improved reconstruction performance compared to individual-modality representations. Additionally, the versatility of such joint representations may be extended beyond integrating audio and video. For example, further joint representations can incorporate textual data and/or sensor inputs. In some cases, the joint representations effectively exploit inter-modal redundancies, thereby making them suitable for multi-modal signal compression applications.

Neural Fields

[0041] Neural Fields are coordinate-based neural networks that represent a signal or a “field.” In some examples, neural fields utilize fully connected neural networks to encode continuous signals, across any dimensions and resolutions, with an adequate set of parameters. In some examples, a Neural Field allows an entire signal to be reconstructed by merely sampling coordinates and feeding them into the corresponding neural network. Applications of such representations can be found in various visual computing problems, including 3D scene reconstruction, 3D audio synthesis, and volumetric video.

[0042] FIG. 1 is a block diagram illustrating a training arrangement (100) that can be used to parameterize a signal using Neural Fields according to some embodiments. The training arrangement (100) includes a neural field component (130) configured to generate a reconstructed signal (140) in response to receiving an input of a positionally encoded coordinate (122). Example architectures of the neural field component (130) are described in more detail below. The positionally encoded coordinate (122) is generated by positionally encoding (120) a specified coordinate (110). In a typical example, a goal of the training the neural field component (130) in the training arrangement (100) is to select various parameters thereof such that a loss function (150) is approximately minimized. Iterative adjustments of the parameters of the neural field component (130) are performed using a backpropagation path (152) that is used to determine the magnitudes of the adjustments based on the computed loss function (150) until the stoppage criteria are met.

[0043] FIG. 2 is a block diagram illustrating an MLP (200) that can be used in the neural field component (130) according to one example. In the example shown, the MLP (200) has four layers (210₁-210₄). The first layer (210₁) is the input layer. The next two layers (210₂, 210₃) are the hidden layers. The fourth layer (210₄) is the output layer. In general, MLP (200) can have N hidden layers, where N is a positive integer. Thus, the specific example of the MLP (200) illustrated in FIG. 2 corresponds to N=2. In some specific examples, the number N is in the range from 1 to 4.

[0044] Each of the layers (210₁-210₄) has M processing elements (PEs) (202), where M is a positive integer. In some specific examples, the number M is in the range from 16 to 256. Each of the PEs (202₂₁-202_{2M}) of the second layer (210₂) is directly connected to receive respective inputs from each of the PEs (202₁₁-202_{1M}) of the input layer (210₁). Each of the PEs (202₃₁-202_{3M}) of the third layer (210₃) is directly connected to receive respective inputs from each of the PEs (202₂₁-202_{2M}) of the second layer (210₂). Each of the PEs (202₄₁-202_{4M}) of the output layer (210₄) is

directly connected to receive respective inputs from each of the PEs (202₃₁-202_{3M}) of the third layer (210₃). Similar PE/layer connections are implemented for additional embodiments corresponding to other values of N. In some examples, each of the PEs (202) operates using a ReLU (Rectified Linear Unit) activation function. In other examples, other suitable activation functions can also be used.

[0045] An entry w_{ab} of the weight matrix $\mathbf{W}$ of the MLP (200) is the weight for the connection between the a -th PE of the subsequent layer and the b -th PE of the previous layer. For a layer having a plurality of PEs, the output $\mathbf{O}$ thereof can be expressed as a function f of the weighted sum of the input values $\mathbf{I}$. The function f is the above-mentioned activation function. In mathematical terms,

$$\mathbf{O} = f(\mathbf{W}\mathbf{I} + \mathbf{B}) \quad (1)$$

where the vector $\mathbf{B}$ is the bias vector. The ReLU is a nonlimiting example of the activation function f and is expressed as follows:

$$f(x) = \begin{cases} x & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \quad (2)$$

[0046] In some examples, one or more of the final MLP outputs are configured to be in the range [0,1]. In such examples, the output layer, e.g., the output layer (210₄), is configured to apply a sigmoid function $S(x)$ to map a wider range of values onto the range [0,1]. The sigmoid function $S(x)$ is expressed as follows:

$$S(x) = \frac{1}{1+e^{-x}} \quad (3)$$

[0047] By using the notation in which the weight matrix $\mathbf{W}$ is expressed as $\{\mathbf{W}_k\}$ and the bias vector $\mathbf{B}$ is expressed $\{\mathbf{b}_k\}$, where the index k denotes the k -th layer of the MLP, and then denoting these parameters as $\Phi = \{\{\mathbf{W}_k\}, \{\mathbf{b}_k\}\}$, one can express the operations performed by the MLP as:

$$\hat{\mathbf{y}} = \text{MLP}_{\Phi}(\mathbf{x}) \quad (4)$$

where $\hat{\mathbf{y}}$ and $\mathbf{x}$ denote the MLP's output and input signals, respectively. With a ground truth signal $\mathbf{y}^{gt}$, the mathematical problem of training the MLP can be formulated as follows:

$$\Phi^* = \arg \min_{\Phi} D(\text{MLP}_{\Phi}(\hat{\mathbf{y}}), \mathbf{y}^{gt}) \quad (5)$$

where $D(\cdot)$ is the loss function (150). A lower value of the loss function $D(\cdot)$ implies lower relative distortion, thereby representing better quality of reconstruction.

[0048] FIG. 3 is a block diagram illustrating a Convolutional Neural Network (CNN) (300) that can be used in the neural field component (130) according to another example. CNNs are beneficial in that they can provide superior performance with image, speech, or audio signal inputs. An example CNN, such as the CNN (300) employs three main types of layers, which are: (i) convolutional layer; (ii) pooling layer; and (iii) fully connected (FC) layer. A convolutional layer is configured to use a filter or Kernel method to extract features. A primary aim of a pooling layer is to decrease the size of the convolved feature map to reduce computational costs. This decrease is accomplished by decreasing the connections between layers and independently operating on each feature map. An FC layer and is used to connect the PEs between two different layers and may be placed before the output layer a CNN architecture. In some examples, a CNN may employ a dropout layer. A dropout layer is a mask that nullifies the contribution of some PEs towards the next layer and leaves unmodified all others.

[0049] FIG. 4 is a block diagram illustrating operations performed in a convolutional layer of the CNN (300) according to an example. In the example shown, a 4x4 image (410) is filtered using a 2x2 kernel (420) to generate a 3x3 activation map (430). The computations performed in the convolutional layer are indicated by the respective formulas shown in the elements of the activation map (430).

[0050] More generally, the operations can be described as a dot product is between two matrices, wherein one matrix represents the set of learnable parameters, such as the kernel (420), and the other matrix represents a portion of the input image or tensor, such as the image (410). In some examples, the convolutional layer has multiple kernels that are spatially smaller than the image but have greater depth. In the forward pass, the kernels move across the height and width of the image, generating a two-dimensional activation map, such as the activation map (430), which indicates the kernel-based filtering at each spatial position in the image. The distance the kernels move during this process is referred to as the stride.

[0051] Let us assume an input of size $W_{in} * W_{in} * C$, where $W_{in} * W_{in}$ are the dimensions of the input tensor and C_{in} is the number of channels of the input tensor. Then, if the layer applies C_{out} number of kernels, each with a spatial size of F and a stride of S , and considering the amount of padding P applied during the operation, the size of the output tensor will be $W_{out} * W_{out} * C_{out}$, where W_{out} can be calculated using the following formula:

$$W_{out} = 1 + \frac{W-F+2P}{s} \quad (6)$$

[0052] In various examples, depending on the type of signal to be parameterized by a Neural Field, the architecture (130) may include different combinations of MLPs, CNNs configured to use various activation functions.

5 [0053] Applicable to all embodiments throughout the specification, instead of inputting a given coordinate directly into a Neural Field, a higher-dimensional embedding may be created from the coordinate. The given coordinate may be a time stamp corresponding to a segment in an audio signal and/or a video frame of a video signal. Since some neural networks may intrinsically be biased towards preferentially learning lower frequency functions, the inputs thereof may be generated by
 10 mapping the initial low-dimensional inputs to a higher dimensional space, for example, by using a series of trigonometric functions applied to the given coordinate, represented as vector γ , for better fitting the output data with high-frequency components. In various examples, the series, represented as vector γ , acting on a coordinate p is defined as follows:

$$\gamma(p) = [\sin(2^{l_0}\pi p) \cos(2^{l_0}\pi p) \sin(2^{l_1}\pi p) \cos(2^{l_1}\pi p) \dots \sin(2^{l_{L-1}}\pi p) \cos(2^{l_{L-1}}\pi p)] \quad (6a)$$

15 where $2L$ is the number of trigonometric components of γ ; and $\{l_0, l_1, \dots, l_{L-1}\}$ are integers. In some examples, $l_n = n$. Applicable to all embodiments that are based on processing time stamps corresponding to a segment in an audio signal and/or a video frame of a video signal, the time stamp may be positionally encoded such that the positionally encoded time stamp serves as input to the respective Neural Field. As used throughout the specification, $\gamma(t)$ refers to a positionally encoded
 20 time stamp corresponding to a segment in an audio signal and/or a video frame of a video signal. In preferred embodiments, the positional encoding may imply a mapping to a higher dimensional space as disclosed above, for example, by using a series of trigonometric functions γ for better fitting the output data of the Neural Field with high-frequency components, and in some embodiments, according to equation (6a) with p representing the time stamp.

25 [0054] The choice of activation functions used in Neural Fields depends on the type of signal being parameterized. For example, for images and videos the ReLU and SiLU/Swish activation functions perform well. The ReLU activation function is expressed by Eq. (2), as already explained above. SiLU/Swish activation function is expressed as follows:

$$SiLU/Swish: f(x) = x \cdot \left(\frac{1}{1+e^{-x}} \right) \quad (7)$$

In the case of audio signals, the Sine activation function helps parameterize the signal effectively without the need of higher dimensional embeddings on the coordinates. The expression for the Sine activation function is as follows:

$$\text{Sine: } f(x) = \sin(x) \quad (8)$$

5 [0055] Similar to activation functions, the loss functions (150) used to train the Neural Fields depend on the type of signal being reconstructed. In terms of image/video reconstruction, the most the L1 loss, the L2 loss, and the Structural Similarity Index (SSIM) loss can be used. Given a ground truth signal sig_{true} and a reconstructed/predicted signal from the Neural Field sig_{pred} , these loss functions can be expressed as follows:

$$10 \quad \text{L1 loss} = |sig_{true} - sig_{pred}| \quad (9)$$

$$\text{L2 loss} = (sig_{true} - sig_{pred})^2 \quad (10)$$

$$\text{SSIM loss} = l(sig_{true}, sig_{pred}) * c(sig_{true}, sig_{pred}) * s(sig_{true}, sig_{pred}) \quad (11)$$

where $l(sig_{true}, sig_{pred})$ is the luminance similarity; $c(sig_{true}, sig_{pred})$ is the contrast similarity, and $s(sig_{true}, sig_{pred})$ is the structural similarity, which are computed as follows:

$$15 \quad l(sig_{true}, sig_{pred}) = \frac{(2 * \mu_{true} * \mu_{pred}) + C_1}{\mu_{true}^2 + \mu_{pred}^2 + C_1} \quad (12)$$

$$c(sig_{true}, sig_{pred}) = \frac{(2 * \sigma_{true} * \sigma_{pred}) + C_2}{\sigma_{true}^2 + \sigma_{pred}^2 + C_2} \quad (13)$$

$$s(sig_{true}, sig_{pred}) = \frac{\sigma_{true, pred} + C_3}{\sigma_{true} \sigma_{pred} + C_3} \quad (14)$$

where μ_{true}, μ_{pred} are the mean intensities of the original and reconstructed signals, respectively; $\sigma_{true}, \sigma_{pred}$ are the standard deviations of the intensities of the original and reconstructed signals, respectively; and $\sigma_{true, pred}$ is the correlation between the two signals.

[0056] In the case of audio signals, the L1 and L2 losses can be used as well. Another choice of the loss function when dealing with reconstruction of audio spectrograms is the Spectral Convergence Loss, which is expressed as follows:

$$\text{Spectral Convergence Loss} = \frac{\|STFT(sig_{pred}) - STFT(sig_{true})\|_E}{\|STFT(sig_{pred})\|_E} \quad (15)$$

25 where STFT(.) is the Short-Time Fourier Transform of the audio signal, which is explained in more detail in another subsection of this specification.

[0057] To evaluate how well a Neural Field has represented a given signal, various evaluation metrics can be used, depending upon the signal. For example, for images/video, one can use the Peak-Signal-to-Noise ratio (PSNR), which is the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. Because many signals have a very wide dynamic range, PSNR is conveniently expressed as a logarithmic quantity using the decibel scale, and a higher PSNR represents a better quality of the signal. Given a noise-free $m \times n$ image I and its noisy approximation K , the PSNR is defined as:

$$PSNR = 20 \cdot \log_{10} \frac{Max_I}{\sqrt{MSE}}, \quad (16)$$

where Max_I is the maximum possible codeword value by given a bit depth; and MSE is given as:

$$MSE = \frac{1}{m \cdot n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I(i, j) - K(i, j)]^2 \quad (17)$$

[0058] For audio signals, one can utilize the Frechet-Audio-Distance (FAD). The FAD evaluation metric is specifically designed for general audio signals. Unlike some other evaluation metrics for audio signals, FAD tends to more-accurately gauge the perceived effects of various distortions.

[0059] FIG. 5 is a flowchart illustrating a FAD workflow (500) according to one embodiment. Given a model (510) that generates a reconstructed audio signal (512) from a noisy audio signal (502) and further given an original audio signal (504), we first compute embeddings (522, 524) of the audio signal using an audio embeddings model (520). In some examples, the audio embeddings model (520) can be a VGG-like model or a PANN model. For example, the audio signal can be passed to a VGG network, and intermediate features can be extracted for the audio embedding. Next, multivariate Gaussians are computed on the reconstructed audio embeddings $N_{recon}(\mu_{recon}, \sigma_{recon})$ and original audio embeddings $N_{og}(\mu_{og}, \sigma_{og})$. Finally, the FAD (530) is computed as follows:

$$FAD(N_{recon}, N_{og}) = \left\| \mu_{recon} - \mu_{og} \right\|^2 + tr(\sigma_{recon} + \sigma_{og} - 2 \sqrt{\sigma_{recon} \cdot \sigma_{og}}) \quad (18)$$

where tr is the trace of the matrix. The lower the FAD (530), the closer the reconstructed signal (512) is to the original signal (504).

Neural Field Representations for Different Modalities

[0060] When dealing with Neural Field representations for images, one approach is to parameterize a Red-Green-Blue (RGB) image using an MLP based Neural Field, where ϕ represents the parameters of the Neural Field MLP that need to be optimized. For example, we train an MLP (200) that maps pixel coordinates to the R,G,B values of the pixel:

$$(R, G, B) = MLP_{\phi}(x, y) \quad (19)$$

[0061] In some examples, we represent images with Neural Field MLPs using several different approaches. For each approach, we use a Neural Field MLP with the following architecture: an input layer (e.g., 210₁), three hidden layers (e.g., 210₂-210₄), each having 256 nodes, and an output layer (not explicitly shown in FIG. 2) that has three nodes trained to fit an image of the size 512x512x3 ($\sim 2 * 10^5$ parameters). The MLP is trained for approximately 5000 epochs using the L2 loss function and the Adam optimizer. The reconstruction image quality metric is PSNR. Example approaches are as follows:

- Directly training the MLP on pixel coordinates (x, y) ;
- Applying Positional Encoding to the pixel coordinates (x, y) and applying the positionally encoded coordinates to the MLP;
- Using different activation functions, e.g., including the above-mentioned activation functions.

By comparing the experimental results obtained with different approaches, we see that the use of Sine activation functions and/or positional encodings provides a suitable reconstruction of images using Neural Field MLPs.

[0062] FIG. 6 is a block diagram illustrating a Neural Field architecture (600) used for providing Neural Field representations for videos according to one example. With the architecture (600), the approach is to parameterize an entire video (602) with frames of dimension $H * W * 3$ using a combination of an MLP (630) and a CNN (640). Here, we train the MLP (630) that maps time-stamp coordinates (610) to the entire frame of dimension $H * W * 3$ of the video at the corresponding timestamp.

$$(Frame^{H*W*3})_t = NeRV_{\theta}(MLP\phi(t)) \quad (20)$$

where ϕ and θ represent the parameters of the MLP (630) and the CNN (640) respectively to be optimized.

[0063] In one example, to represent the video (602), we consider a NeRV model of size $\sim 3.02 * 10^6$ parameters which is trained to fit a 10 second, 10 fps video fragment where each frame has a dimension of $512 \times 512 \times 3$. The NeRV model is trained for 5000 epochs using the L2 loss function and the Adam optimizer. In some examples, we alternatively train the NeRV models with the SSIM loss or a combination of L2/L1 loss and SSIM loss. The reconstruction video quality metric is PSNR.

[0064] When dealing with Neural Field representations for audio, one approach is to directly represent the amplitude samples at each time index of the audio signal with an MLP. Here, we represent the audio signal directly in terms of its amplitude samples as follows:

$$Amplitude_a = MLP_{\phi}(a) \quad (21)$$

where a is the amplitude index and ϕ represents the parameters of the MLP to be optimized.

[0065] In one example, we reconstruct a music signal of a 5-second duration with a sample rate of 16 kHz. In this example, the total number of amplitude indices is 80000. For the representation, we use a Neural Field MLP (200) with the following architecture: an input layer and five hidden layers with 256 PEs (202) each and an output layer with one PE ($\sim 6 * 10^5$ parameters). We utilize Sine activations (see Eq. (8)). The MLP (200) is trained for 5000 epochs, and the MLP parameters are optimized using the L2 loss and the Adam optimizer. The training workflow is the FAD workflow (500).

[0066] When dealing with a 3D scene representation, one approach is to use the Neural Radiance Fields (NeRFs). A NeRF is a fully connected neural network that can generate novel views of complex 3D scenes, based on a partial set of 2D images. The NeRF model takes in the spatial coordinates (x, y, z) and viewing directions (θ, ϵ) , where θ and ϵ are the polar coordinates of the direction from a query point to the camera. These five coordinates are taken as an input to generate outputs of volume density and view-dependent emitted radiance for each coordinate. By employing suitable volume rendering, a NeRF is used to construct 3D scene representations from 2D images, e.g., as follows:

$$NeRF: (R, G, B, \sigma) = MLP_{\phi}(x, y, z, \theta, \epsilon) \quad (22)$$

where ϕ represents the parameters of the NeRF MLP to be optimized.

[0067] One application of Neural Fields in the audio domain is the creation of Neural Acoustic Fields (NAFs), which are designed to capture the propagation of sound in a physical environment.

NAFs represent spatial reverberation in the physical environment based on an impulse response.

This impulse response is modeled as a function dependent on various factors, including the emitter location (x, y) , the listener location (x', y') , frequency (f) , time (t) , orientation θ , and the left/right ear index (k) . By treating acoustic propagation in the scene as a linear time-invariant system, NAFs
 5 (Ω_{grid}) learn to continuously map all emitter and listener location pairs to a neural impulse response function, which can then be applied to any arbitrary sounds. The continuous nature of NAFs enables the rendering of spatial acoustics for a listener at any location and facilitates sound propagation prediction at new locations, e.g., as follows:

$$NAF: [v_{\{stft_{mag}\}}(t, f), v_{\{stft_{phase}\}}(t, f)] = \Omega_{\emptyset}(x, y, \theta, k, x', y', t, f) \quad (23)$$

10 where $v_{\{stft_{mag}\}}(t, f)$ and $v_{\{stft_{phase}\}}(t, f)$ are the NAF predicted magnitude and phase spectrograms and $\emptyset$ represents the weights of the NAF Ω to be optimized.

[0068] When dealing with Volumetric Videos, one approach is to use NeuVV utilized Neural Fields to generate immersive, interactive, and spatial-temporal rendering of volumetric performances with photo-realism and in real-time. A NeuVV samples 4D points (x, y, z, t) within
 15 the volumetric video and then utilizes an MLP-based neural module that predicts coefficients of 4D Hyperspherical Harmonics and a hyper angle. Using these predictions, the 3D spherical Harmonics are sliced to predict density σ and view-dependent color c thereby allowing real-time volumetric video rendering and editing, e.g., as follows:

$$NeuVV: (c, \sigma) = MLP_{\phi}(x, y, z, t) \quad (24)$$

20 where ϕ represents the parameters of the NeUVV MLP to be optimized.

Example Embodiments for Improved Neural Field Representations of Audio Signals

[0069] In this subsection, we describe embodiments for the representation of audio signals using Neural Fields based on two approaches:

- Short-Time Fourier Transform (STFT) Magnitude and Phase spectrograms
 25 representation of audio signals; and
- Time stamp based Neural Field Representations of Audio Signals.

[0070] Spectrograms of audio signals are image-like representations that can be used for audio reconstruction use cases. In some examples, we recreate the STFT magnitude and phase

spectrograms directly allowing us to revert to the original audio signal without losing phase information.

[0071] FIGS. 7A-7B graphically illustrate window functions used in STFT according to some examples. More specifically, FIG. 7A graphically illustrates a Hamming window function (702).

5 FIG. 7B graphically illustrates a rectangular window function (704).

[0072] Unlike commonly used frequency-domain representations, such as the Discrete-time Fourier Transform (DTFT) and Discrete Fourier Transform (DFT), which provide averaged frequency components over the signal duration, the STFT focuses on smaller segments of the signal by applying a Fourier transform to each short-duration segment. This process involves multiplying
10 a longer time function $x[n]$ by a shorter window function $w[n]$, examples of which are shown in FIGS. 7A-7B. The rectangular window function (704) operates to directly extract the desired short sequence without any further modifications. The Hamming window function (702) operates to apply a taper to the ends, thereby tending to enhance the representation in the frequency domain.

[0073] For a signal $x[n]$ ($-\infty$ to $+\infty$), and a window function $w[n]$, the STFT can be described
15 as follows:

$$X[n, w] = \sum_{m=-\infty}^{\infty} w[n-m] x[m] e^{-jwm} \quad (25)$$

In some practical applications, it is useful to assess the STFT at a finite set of equally spaced points along the frequency axis, similar to how the DFT is often preferred over the DTFT in conventional digital signal processing tasks. To denote the number of discrete frequency channels utilized in the
20 STFT, we will use the variable N , while the variable N_w will indicate the length of the finite-duration window function $w[n]$. Under these notational conventions, the frequencies for which the STFT is computed are given by $w_k = 2\pi k/N$. When employing a finite-duration window with values of n ranging from 0 to $N_w - 1$, the STFT equation takes the following form:

$$X[n, k] = \sum_{m=n-(N_w-1)}^n w[n-m] x[m] e^{-jw_k m} = \sum_{m=n-(N_w-1)}^n w[n-m] x[m] e^{-j2\pi m k/N} \quad (26)$$

25

where $X[n, k]$ is a function of both time and frequency and now both the time and frequency variables are discrete. The variable n denotes the location of the analysis window along the time axis, and the segment of time delimited by the window is referred to as the analysis frame. The variable k is a frequency index and is referred to as a frequency bin. One can think of the STFT as
30 representing the DFT of the finite-duration time function $x[m]w[n-m]$. Here, the variable m is a

“dummy” time argument, and the variable n identifies the location of the short segment of the original time function as it is extracted using the window $w[n - m]$, which moves along the $m - axis$ according to the value of n .

[0074] FIG. 8 is a block diagram illustrating a workflow (800) for generating spectrograms via the STFT according to one example. Therein, an inset illustrates an audio signal (804) that is being processed. The audio signal (804) is partitioned into a sequence of overlapping segments (806₀-806_M). For illustration purposes, only six of such segments (806₀-806₅) are explicitly shown. A selected window function $w[n]$ (e.g., 702 or 704) is then applied to each of the segments (806₀-806_M), and each resulting window-function scaled segment is subjected to a DFT (810) to generate a respective complex-valued spectrum (812₀-812_M). For illustration purposes, only six of such spectra (812₀-812₅) are explicitly shown. Each of the spectra (812₀-812_M) is then split into a magnitude component and a phase component, and the components are arranged via a stacking operation (820) into a magnitude spectrogram (822) and a phase spectrogram (824).

[0075] It should be noted that the STFT is perfectly invertible, meaning that a complex valued STFT can be converted back to the original waveform. This inverse operation is referred to as the inverse Short-time Fourier Transform or iSTFT. For example, the STFT spectrograms (822, 824) can be subjected to the iSTFT to reconstruct the audio signal (804).

[0076] FIG. 9 is a block diagram illustrating a neural-fields-based representation (900) of an audio signal using the corresponding STFT magnitude and phase spectrograms according to one embodiment. The representation (900) employs an MLP (910) with the PEs thereof configured to use the Sine activation function. An input to the MLP (910) are the (x, y) coordinates ($MagSpec_{x,y}, PhaseSpec_{x,y}$) for the magnitude and phase spectrograms encoded in the MLP. An output of the MLP (910) are the spectrogram values $MagSpec_{value}, PhaseSpec_{value}$ corresponding to the (x, y) coordinates. Eq. (27) provides a mathematical expression describing the representation (900):

$$MagSpec_{value}, PhaseSpec_{value} = MLP_{\phi}(MagSpec_{x,y}, PhaseSpec_{x,y}) \quad (27)$$

where ϕ are the parameters of the MLP that determined via a training process (e.g., see FIG. 5).

[0077] FIGS. 10A-10C graphically illustrate audio signal reconstruction results obtained using the neural-fields-based representation (900) according to one example. In FIGS. 10A-10C, the left graphs (1002, 1006, 1010) show the magnitude spectrogram (e.g., analogous to the magnitude

spectrogram (822), FIG. 8), the phase spectrogram (e.g., analogous to phase spectrogram (824), FIG. 8), and the audio signal waveform (e.g., analogous to the audio signal (804), FIG. 8), respectively, of an original music signal having a five second duration with a sampling rate of 16 kHz. The right graphs (1004, 1008, 1012) show the magnitude spectrogram, the phase spectrogram, and the audio signal waveform of a reconstructed audio signal, respectively, obtained using the neural-fields-based representation (900).

[0078] In this case, we take an STFT of the original music signal which results in the Magnitude Spectrogram (1002) (log scale) and the Phase Spectrogram (1006) which are stacked and zero padded to an equal dimension. For the representation (900), we use an MLP (200) having the size $\sim 6 * 10^5$ parameters, which has 10 hidden layers of the size 256 each and uses Sine activations in the PEs thereof. The MLP (200) is trained for 5000 epochs and the network parameters are optimized using the L2 loss and the Adam optimizer. A reconstructed audio signal is obtained by computing the iSTFT of the reconstructed magnitude and phase spectrograms. Then, we compute the FAD between the original signal and the reconstructed signal to see how well the representation works (also see FIG. 5). One benefit of this approach is its ability to provide image-like representations of audio signals.

[0079] The time stamp based Neural Field Representations of Audio Signals disclosed herein address certain challenges associated integrating an audio modality with other modalities, such as the video modality. Some of such challenges come from the difficulties encountered when associating audio samples with video frames. Beneficially, the time-stamp-based Neural Field representation for audio signals facilitates such association of audio with other modalities, including video. For instance, when we have an audio signal sampled at 16 kHz and a video signal with a frame rate of 10 frames per second (fps), each frame can be linked with 1600 audio samples. This method allows us to represent a group of audio samples together at a specific time stamp instead of individually representing each sample with an amplitude index or combining separate spectrograms. This attribute enables seamless integration and synchronization of audio with other multimodal sources.

[0080] Example embodiments disclosed herein use two types of Neural Field representations to represent and reconstruct audio samples in a time-stamp manner:

(1) Using MLPs and Sine Activations; and

(2) Using MLPs along with CNNs with Sine Activations (NeRA).

These two types of Neural Field representations are discussed in more detail below in reference to FIGS. 11-16.

[0081] FIG. 11 is a block diagram of a time-stamp-based representation (1100) of an audio signal according to one embodiment. The representation (1100) utilizes both positional encodings $\gamma(t)$ (1104) and an MLP-based network (1110) with sine activations to reconstruct audio segments (1120) at different timestamps (t) (1102). Eq. (28) provides a mathematical expression describing the representation (1100):

$$AudioSamples_t = MLP_{\phi}(\gamma(t)) \quad (28)$$

where ϕ represents the parameters of the MLP-based network (1110).

[0082] In one example, the MLP-based network (1110) is trained to represent an audio signal of 5 s in duration and at 16 kHz sampling rate with both speech and music present therein. The size of the Neural Field is characterized by $\sim 6 * 10^5$ parameters and has 10 hidden layers of size 256 each. The neural field has been trained for 5000 epochs. The weights have been optimized using the L2 loss function and Adam optimizer.

[0083] FIGS. 12A-12B graphically illustrate audio signal reconstruction results obtained using the neural-fields-based representation (1100) according to one example. FIG. 12A illustrates an original audio signal (1202). FIG. 12B illustrates a reconstructed audio signal (1204), which is characterized by the indicated FAD score. The shown results indicate that the use of both positional encodings and Sine activations can lead to an improved quality of the audio signal representation by a Neural Field MLP.

[0084] FIG. 13 is a block diagram of a time-stamp-based representation (1300) of an audio signal according to another embodiment. The time-stamp-based representation (1300) includes an MLP (1310) and a CNN (1316) serially connected to one another as indicated in FIG. 13. The CNN (1316) includes a plurality of NeRA blocks (1318), which are described in more detail below in reference to FIG. 14. In the example shown, the activation functions used in the MLP (1310) and a CNN (1316) are the Sine activation functions (also see Eq. (8)). The representation (1300) also utilizes positional encodings $\gamma(t)$ (1304) of timestamps (t) (1302) prior to the MLP (1310). In operation, the representation (1300) reconstructs audio segments (1320) in response to different

timestamps (t) (1302). Eq. (29) provides a mathematical expression describing the representation (1300):

$$\text{AudioSamples}_t = \text{NeRA}_\theta (\text{MLP}_\phi(\gamma(t))) \quad (29)$$

where θ and ϕ represent the parameters of the CNN (1316) and the MLP (1310), respectively.

5 [0085] FIG. 14 is a block diagram illustrating a NeRA block (1318) used in the CNN (1316) according to one example. In the example shown, the NeRA block (1318) includes a convolution layer (1402), a sub-pixel convolution layer (1404), and a sine activation subblock (1406). Example operations of the convolution layer (1402) have been described above in reference to FIG. 4. The sub-pixel convolution layer (1404) performs upsampling operations on the output of the convolution
10 layer (1402). The sine activation subblock (1406) applies the Sine activation functions in accordance with Eq. (8).

[0086] In one example, the time-stamp-based representation (1300) is trained to represent an audio signal of 5 s in duration and at 16 kHz sampling rate with both speech and music present therein. The size of the Neural Field is characterized by $\sim 6 * 10^5$ parameters, including the
15 parameters of the MLP (1300) with five hidden layers of size 256 and the parameters of the three NeRA blocks (1318) of the CNN (1316). The neural field has been trained for 5000 epochs. The weights have been optimized using the L2 loss and Adam optimizer.

[0087] FIGS. 15A-15B graphically illustrate audio signal reconstruction results obtained using the time-stamp-based representation (1300) according to one example. FIG. 15A illustrates an
20 original audio signal (1502). FIG. 15B illustrates a reconstructed audio signal (1504), which is characterized by the indicated FAD score. The shown results indicate that the combined use of positional encodings, MLP+CNN neural fields, and Sine activations can lead to a further improved quality of audio signal representations.

[0088] FIG. 16 is a table comparing the performance of the time-stamp-based representations
25 (1100, 1300) according to various examples. More specifically, the table shows the FAD scores for different audio signals, such as rain, speech, etc. The compared models (1100, 1300) have the same number of trainable parameters. From the FAD scores shown in the table, one can see that the lowest FAD score is 0.003. Although the NeRA (1300) appears to outperform the MLP (1100) in terms of the FAD scores, the FAD score values are relatively low in both cases, causing the
30 difference in the quality of the reconstructed audio to be practically imperceptible to the listener.

Example Embodiments for Joint Neural Field Representation of Video and Audio (NeRVA)

[0089] FIG. 17 is a block diagram illustrating a joint Neural Field representation (1700) of video and audio signals according to one embodiment. The NeRVA representation (1700) effectively represents both audio and video signals together, allowing for time-stamp-based reconstruction of those signals together. In the example shown, the NeRVA representation (1700) includes the following components:

1. A Positional Encoding block (1704);
2. A Modality-Unifying Audio-Video MLP (1710), with Sine activations;
3. A reshaping MLP block (1720); and
4. A reconstruction CNN block (1740).

[0090] Eq. (30) provides a mathematical expression describing the NeRVA representation (1700):

$$(VideoFrame(t)_{r,g,b}, AudioSample(t)_{Amplitude}) = NeRVA_{\tau}(t) \quad (30)$$

where t is the timestamp coordinate; and τ represents the (optimizable) parameters of the entire

NeRVA representation (1700). Eq. (30) can further be expanded as follows:

$$NeRVA_{\tau}(t) = VideoRecon_{\theta}(AVMLP_{\phi}(t)) + AudioRecon_{\omega}(AVMLP_{\phi}(t)) \quad (31)$$

where θ and ω represent the (optimizable) parameters of a video reconstruction branch (1722, 1742) and an audio reconstruction branch (1724, 1744) of the NeRVA representation (1700); and ϕ represents the (optimizable) parameters of the Audio-Video MLP (1710) "AVMLP". The video and audio reconstruction branches can further be represented as follows:

$$VideoRecon_{\theta}(AVMLP_{\phi}(t)) = NeRV_{\beta}(VideoStem_{\alpha}(AVMLP_{\phi}(\gamma(t)))) \quad (32)$$

where α and β are the parameters of a Video Stem MLP (1722) in the reshaping MLP block (1720) and of NeRV blocks (1742) in the reconstruction CNN block (1740), respectively; and

$$AudioRecon_{\omega}(AVMLP_{\phi}(t)) = NeRA_{\epsilon}(AudioStem_{\tau}(AVMLP_{\phi}(\gamma(t)))) \quad (33)$$

where ϵ and τ are the parameters of an Audio Stem MLP (1724) in the reshaping MLP block (1720) and of NeRA blocks (1744) in the reconstruction CNN block (1740), respectively.

[0091] The Positional Encoding block (1704) employs the positional encodings described above (e.g., see Eq. (6a)) to create higher-dimensional embeddings $\gamma(t)$ (1706) of input time-stamp coordinates t (1702). The embeddings $\gamma(t)$ (1706) are then passed as inputs to the Audio-Video MLP (1710). The Audio-Video MLP (1710) processes the embeddings $\gamma(t)$ (1706) to unify the

audio and video modalities, resulting in an Audio-Video (AV) embedding $AVMLP_{\phi}(\gamma(t))$ (1712). The parameters of the Audio-Video MLP (1710) are learned jointly, i.e., both the audio and video losses are backpropagated through this MLP. In one example, the Audio-Video MLP (1710) is configured to utilize Sine activations between the MLP layers. The AV embedding $AVMLP_{\phi}(\gamma(t))$ (1712) generated by the Audio-Video MLP (1710) is passed onto the video reconstruction branch (1722, 1742) and the audio reconstruction branch (1724, 1744).

[0092] The Video Stem MLP (1722) and the Audio Stem MLP (1724) operate to reshape the AV embedding (1712) into a respective 3D tensor (1732, 1734). Each of these stem MLPs includes one fully connected layer configured to create a 1D embedding that is then reshaped into the respective 3D tensor. The tensor (1732) has the size of C^V, H^V, W^V and is directed to the NeRV blocks (1742). The tensor (1734) has the size of C^A, H^A, W^A and is directed to the NeRA blocks (1744). Each of the stem MLPs (1722, 1724) is trained only with the loss of the reconstruction branch it is a part of. In some examples, the stem MLPs (1722, 1724) may be absent in order to reduce the total number of parameters in the NeRVA architecture.

[0093] The tensor output (1732) of the video stem MLP (1722) is passed onto the NeRV blocks (1742) referred to as first CNN. In one example, an individual NeRV block has a structure similar to that shown in FIG. 14 but employs Swish activation (Eq. (7)) instead of Sine activation. The tensor output (1734) of the audio stem MLP (1724) is passed onto the NeRA blocks (1744) referred to as second CNN. In one example, an individual NeRA block has a structure shown in FIG. 14, with Sine activations. Typically, the number of the NeRA blocks (1744) is smaller than the number of the NeRV blocks (1742) due to a difference in the complexity of audio segments and RGB frames. The outputs of the NeRV blocks (1742) and the NeRA blocks (1744) are a reconstructed video frame (1750) and a reconstructed audio segment (1760), respectively, corresponding to the input time-stamp coordinates (1702).

[0094] A notable feature of the architecture illustrated in FIG. 17 is its straightforward upgradability to incorporate more reconstruction branches, which can be used for seamless unification of more than two modalities within the same framework. In some examples, the joint Neural Field representation (1700) can be trained with the L2 Loss function being applied to both audio and video reconstruction branches. In other examples, different respective loss functions can be used for training the audio and video reconstruction branches. For example, in some cases, the

SSIM loss can be used for training the video, and the L1 Loss can be used for training the Audio branch.

[0095] FIG. 18 is a block diagram illustrating a joint Neural Field representation (1800) of video and audio signals according to another embodiment. The representation (1800) can conceptually be viewed as a modification of the representation (1700), with the modification being the removal of the reshaping MLP block (1720). All remaining details of the embodiment shown in and described with respect to FIG. 17 also apply to the embodiment shown in FIG. 18. By removing the reshaping MLP block (1720), we effectively reduce the number of parameters in the NeRVA model.

However, this modification also imposes a fixed reshaping of the AV embedding (1712)

(C^{AV}, H^{AV}, W^{AV}), where C^{AV} denotes the channels, H^{AV} denotes the height, and W^{AV} denotes the width of the 3D tensor obtained by reshaping the 1D AV embedding. Because the NeRV blocks (1742), the first CNN, and the NeRA blocks (1744), the second CNN, receive the inputs of the same dimensions, the flexibility in selecting the CNN parameters for the NeRV and NeRA blocks (1742, 1744) is more limited in the representation (1800) than in the representation (1700). While the representation (1800) provides a more streamlined model, the streamlining comes with certain trade-offs in terms of adaptability and parameter tuning for specific modalities.

Example Hardware

[0096] FIG. 19 is a block diagram illustrating a computing device (1900) used to implement neural field representations according to various embodiments. The computing device (1900) comprises input/output (I/O) devices (1910), a processing engine (1920), and a memory (1930). The I/O devices (1910) may be used to enable the device (1900) to receive various input signals (1902) and to output various output signals (1904). For example, the I/O devices (1910) may be operatively connected to send and receive signals via a communication channel.

[0097] The memory (1930) may have buffers to receive data. Once the data are received, the memory (1930) may provide parts of the data to the processing engine (1920) for processing therein. The processing engine (1920) includes a processor (1922) and a memory (1924). The memory (1924) may store therein program code, which when executed by the processor (1922) enables the processing engine (1920) to perform various data processing operations, including but not limited to at least some operations associated with training and using any of the above-described neural field representations.

[0098] According to one example embodiment disclosed above, e.g., in the summary section and/or in reference to any one or any combination of some or all of FIGS. 1-19, provided is a method of representing an audio signal using a neural field, the method comprising: transforming the audio signal into a pair of 2D spectrograms including a 2D magnitude spectrogram and a
5 corresponding 2D phase spectrogram, a first dimension of the 2D spectrograms representing time, a second dimension of the 2D spectrograms representing frequency; and training a neural-field network to represent the pair of 2D spectrograms, the neural-field network being configured to output a magnitude value and a phase value corresponding to the audio signal in response to an input of a 2D coordinate, the training producing a set of fixed parameters for the neural field network.

10 [0099] In some embodiments of the above method, the neural-field network is implemented using a multilayer perceptron (MLP).

[00100] In some embodiments of any of the above methods, processing elements of the MLP are configured to use an activation function selected from the group consisting of a ReLU activation function, a SiLU/Swish activation function, and a Sine activation function.

15 [00101] In some embodiments of any of the above methods, the transforming comprises: dividing the audio signal into a plurality of segments, each of the segments corresponding to a different respective time interval; applying a Fourier transform to each of the segments to generate a corresponding spectrum; splitting the corresponding spectrum into a respective magnitude component and a respective phase component; stacking a plurality of the respective magnitude
20 components along the first dimension to form the 2D magnitude spectrogram; and stacking a plurality of the respective phase components along the first dimension to form the 2D phase spectrogram.

[00102] In some embodiments of any of the above methods, the dividing comprises applying a window function to the audio signal.

25 [00103] In some embodiments of any of the above methods, the plurality of segments includes overlapping segments.

[00104] According to another example embodiment disclosed above, e.g., in the summary section and/or in reference to any one or any combination of some or all of FIGS. 1-19, provided is a method of representing an audio signal using a neural field, the method comprising: generating a

pair of 2D spectrograms including a 2D magnitude spectrogram and a corresponding 2D phase spectrogram by testing a neural-field network that has been trained to output a magnitude value and a phase value corresponding to the audio signal in response to an input of a 2D coordinate, a first dimension of the 2D spectrograms representing time, a second dimension of the 2D spectrograms representing frequency; and reconstructing the audio signal based on the pair of 2D spectrograms.

[00105] In some embodiments of the above method, the neural-field network is implemented using a multilayer perceptron (MLP).

[00106] In some embodiments of any of the above methods, processing elements of the MLP are configured to use an activation function selected from the group consisting of a ReLU activation function, a SiLU/Swish activation function, and a Sine activation function.

[00107] In some embodiments of any of the above methods, the reconstructing comprises: obtaining a plurality of spectra by pairwise combining magnitude components obtained from the 2D magnitude spectrogram and corresponding phase components obtained from the corresponding 2D phase spectrogram; applying an inverse Fourier transform to each of the spectra to generate a corresponding time-domain segment; and assembling a plurality of the corresponding time-segments into the reconstructed audio signal.

[00108] According to yet another example embodiment disclosed above, e.g., in the summary section and/or in reference to any one or any combination of some or all of FIGS. 1-19, provided is a method of representing an audio signal using a neural field, the method comprising: applying positional encoding to a time stamp corresponding to a segment of the audio signal to generate a respective higher-dimensional embedding; and training a neural-field network to represent a plurality of segments of the audio signal, the neural-field network being configured to output a respective set of audio samples representing a respective one of the plurality of segments of the audio signal in response to an input of the respective higher-dimensional embedding, the training producing a set of fixed parameters for the neural field network.

[00109] In some embodiments of the above method, the neural-field network includes a multilayer perceptron (MLP).

[00110] In some embodiments of any of the above methods, processing elements of the MLP are configured to use an activation function selected from the group consisting of a ReLU activation function, a SiLU/Swish activation function, and a Sine activation function.

5 [00111] In some embodiments of any of the above methods, the MLP is serially connected with a convolutional neural network (CNN).

[00112] In some embodiments of any of the above methods, the CNN comprises a plurality of serially connected blocks, each of the blocks including a respective convolutional layer and a respective sub-pixel convolutional layer and having processing elements configured to use a Sine activation function.

10 [00113] According to yet another example embodiment disclosed above, e.g., in the summary section and/or in reference to any one or any combination of some or all of FIGS. 1-19, provided is a method of representing an audio signal using a neural field, the method comprising: applying positional encoding to a time stamp corresponding to a segment of the audio signal to generate a higher-dimensional embedding; and generating a set of audio samples representing the segment of
15 the audio signal by applying the higher-dimensional embedding to a neural-field network trained to represent a plurality of segments of the audio signal, the neural-field network being configured to output a respective set of audio samples representing a respective one of the plurality of segments of the audio signal in response to an input of a respective higher-dimensional embedding.

20 [00114] In some embodiments of the above method, the neural-field network includes a multilayer perceptron (MLP).

[00115] In some embodiments of any of the above methods, processing elements of the MLP are configured to use an activation function selected from the group consisting of a ReLU activation function, a SiLU/Swish activation function, and a Sine activation function.

25 [00116] In some embodiments of any of the above methods, the MLP is serially connected with a convolutional neural network (CNN).

[00117] In some embodiments of any of the above methods, the CNN comprises a plurality of serially connected blocks, each of the blocks including a respective convolutional layer and a

respective sub-pixel convolutional layer and having processing elements configured to use a Sine activation function.

[00118] According to yet another example embodiment disclosed above, e.g., in the summary section and/or in reference to any one or any combination of some or all of FIGS. 1-19, provided is a method of jointly representing an audio signal and a video signal using a neural field, the method comprising: applying positional encoding to a time stamp corresponding to a video frame of the video signal and an associated segment of the audio signal to generate a respective higher-dimensional embedding; and training a neural-field network to represent a sequence of video frames of the video signal and an associated plurality of segments of the audio signal, the neural-field network being configured to output a respective video frame of the sequence and a respective set of audio samples representing a respective one of the associated plurality of segments in response to an input of the respective higher-dimensional embedding, the training producing a set of fixed parameters for the neural field network.

[00119] In some embodiments of the above method, the neural-field network comprises a first multilayer perceptron (MLP) configured to generate a respective joint audio-video embedding in response to the respective higher-dimensional embedding.

[00120] In some embodiments of any of the above methods, the training comprises determining parameters for the first MLP based on backpropagating both audio losses and video losses through the first MLP.

[00121] In some embodiments of any of the above methods, the neural-field network further comprises: a first branch configured to generate the respective video frame in response to the respective joint audio-video embedding; and a second branch configured to generate the respective one of the associated plurality of segments of the audio signal in response to the respective joint audio-video or multimedia embedding.

[00122] In some embodiments of any of the above methods, the first branch further comprises a second MLP coupled between the first MLP and the first CNN; and wherein the second branch further comprises a third MLP coupled between the first MLP and the second CNN.

[00123] In some embodiments of any of the above methods, the first CNN comprises a plurality of serially connected first blocks, each of the first blocks including a respective convolutional layer

and a respective sub-pixel convolutional layer and having processing elements configured to use a SiLU/Swish activation function; and wherein the second CNN comprises a plurality of serially connected second blocks, each of the second blocks including a respective convolutional layer and a respective sub-pixel convolutional layer and having processing elements configured to use a Sine
5 activation function.

[00124] In some embodiments of any of the above methods, a total number of the second blocks is smaller than a total number of the first blocks.

[00125] According to yet another example embodiment disclosed above, e.g., in the summary section and/or in reference to any one or any combination of some or all of FIGS. 1-19, provided is a
10 method of jointly representing an audio signal and a video signal using a neural field, the method comprising: applying positional encoding to a time stamp corresponding to a video frame of the video signal and further corresponding to an associated segment of the audio signal to generate a respective higher-dimensional embedding; and reconstructing the video frame of the video signal and generating a corresponding set of audio samples representing the associated segment of the
15 audio signal by applying the higher-dimensional embedding to a neural-field network trained to represent a sequence of video frames of the video signal and an associated plurality of segments of the audio signal, the neural-field network being configured to output a respective video frame of the sequence and a respective set of audio samples representing a respective one of the associated plurality of segments in response to an input of the respective higher-dimensional embedding, the
20 training producing a set of fixed parameters for the neural field network.

[00126] In some embodiments of the above method, the neural-field network comprises a first multilayer perceptron (MLP) configured to generate a respective joint audio-video embedding in response to the respective higher-dimensional embedding.

[00127] In some embodiments of any of the above methods, the first MLP has been trained based
25 on backpropagating both audio losses and video losses through the first MLP.

[00128] In some embodiments of any of the above methods, the neural-field network further comprises: a first branch configured to generate the respective video frame in response to the respective joint audio-video embedding; and a second branch configured to generate the respective

one of the associated plurality of segments in response to the respective joint audio-video embedding.

[00129] In some embodiments of any of the above methods, the first branch comprises a first convolutional neural network (CNN); and wherein the second branch comprises a second CNN.

- 5 **[00130]** In some embodiments of any of the above methods, the first branch further comprises a second MLP coupled between the first MLP and the first CNN; and wherein the second branch further comprises a third MLP coupled between the first MLP and the second CNN.

- 10 **[00131]** In some embodiments of any of the above methods, the first CNN comprises a plurality of serially connected first blocks, each of the first blocks including a respective convolutional layer and a respective sub-pixel convolutional layer and having processing elements configured to use a SiLU/Swish activation function; and wherein the second CNN comprises a plurality of serially connected second blocks, each of the second blocks including a respective convolutional layer and a respective sub-pixel convolutional layer and having processing elements configured to use a Sine activation function.

- 15 **[00132]** In some embodiments of any of the above methods, a total number of the second blocks is smaller than a total number of the first blocks.

- 20 **[00133]** According to yet another example embodiment disclosed above, e.g., in the summary section and/or in reference to any one or any combination of some or all of FIGS. 1-19, provided is a method of jointly representing a first multimedia signal and a second multimedia signal using a neural field, the method comprising: applying positional encoding to a time stamp corresponding to a portion of the second multimedia signal and an associated segment of the first multimedia signal to generate a respective higher-dimensional embedding; and training a neural-field network to represent a sequence of portions of the second multimedia signal and an associated plurality of segments of the first multimedia signal, the neural-field network being configured to output a
25 respective portion of the sequence and a respective set of samples representing a respective one of the associated plurality of segments in response to an input of the respective higher-dimensional embedding, the training producing a set of fixed parameters for the neural field network.

[00134] In some embodiments of the above method, the neural-field network comprises a first multilayer perceptron (MLP) configured to generate a respective joint multimedia embedding in response to the respective higher-dimensional embedding.

[00135] In some embodiments of any of the above methods, the neural-field network further comprises: a first branch configured to generate the respective portion of the sequence in response to the respective joint multimedia embedding; and a second branch configured to generate the respective one of the associated plurality of segments in response to the respective joint multimedia embedding.

[00136] In some embodiments of any of the above methods, the first branch comprises a first convolutional neural network (CNN); and wherein the second branch comprises a second CNN.

[00137] In some embodiments of any of the above methods, the first branch further comprises a second MLP coupled between the first MLP and the first CNN; and wherein the second branch further comprises a third MLP coupled between the first MLP and the second CNN.

[00138] In some embodiments of any of the above methods, the first multimedia signal is a 1D signal having a first sampling rate; and wherein the second multimedia signal is a 3D signal having a different sampling rate.

[00139] In some embodiments of any of the above methods, the 3D signal includes a respective 1D signal and a respective 2D signal; wherein the respective 2D signal has a second sampling rate that is lower than the first sampling rate; and wherein the respective 1D signal has a third sampling rate that is lower than the second sampling rate.

[00140] According to yet another example embodiment disclosed above, e.g., in the summary section and/or in reference to any one or any combination of some or all of FIGS. 1-19, provided is a non-transitory computer-readable medium storing instructions that, when executed by an electronic processor, cause the electronic processor to perform operations comprising any of the above methods.

[00141] With regard to the processes, systems, methods, heuristics, etc. described herein, it should be understood that, although the steps of such processes, etc. have been described as occurring according to a certain ordered sequence, such processes could be practiced with the

described steps performed in an order other than the order described herein. It further should be understood that certain steps could be performed simultaneously, that other steps could be added, or that certain steps described herein could be omitted. In other words, the descriptions of processes herein are provided for the purpose of illustrating certain embodiments and should in no way be construed so as to limit the claims.

[00142] Accordingly, it is to be understood that the above description is intended to be illustrative and not restrictive. Many embodiments and applications other than the examples provided would be apparent upon reading the above description. The scope should be determined, not with reference to the above description, but should instead be determined with reference to the appended claims, along with the full scope of equivalents to which such claims are entitled. It is anticipated and intended that future developments will occur in the technologies discussed herein, and that the disclosed systems and methods will be incorporated into such future embodiments. In sum, it should be understood that the application is capable of modification and variation.

[00143] All terms used in the claims are intended to be given their broadest reasonable constructions and their ordinary meanings as understood by those knowledgeable in the technologies described herein unless an explicit indication to the contrary is made herein. In particular, use of the singular articles such as “a,” “the,” “said,” etc. should be read to recite one or more of the indicated elements unless a claim recites an explicit limitation to the contrary.

[00144] The Abstract of the Disclosure is provided to allow the reader to quickly ascertain the nature of the technical disclosure. It is submitted with the understanding that it will not be used to interpret or limit the scope or meaning of the claims. In addition, in the foregoing Detailed Description, it can be seen that various features are grouped together in various embodiments for the purpose of streamlining the disclosure. This method of disclosure is not to be interpreted as reflecting an intention that the claimed embodiments incorporate more features than are expressly recited in each claim. Rather, as the following claims reflect, inventive subject matter lies in fewer than all features of a single disclosed embodiment. Thus, the following claims are hereby incorporated into the Detailed Description, with each claim standing on its own as a separately claimed subject matter.

[00145] While this disclosure includes references to illustrative embodiments, this specification is not intended to be construed in a limiting sense. Various modifications of the described

embodiments, as well as other embodiments within the scope of the disclosure, which are apparent to persons skilled in the art to which the disclosure pertains are deemed to lie within the principle and scope of the disclosure, e.g., as expressed in the following claims.

5 [00146] Some embodiments may be implemented as circuit-based processes, including possible implementation on a single integrated circuit.

10 [00147] Some embodiments can be embodied in the form of methods and apparatuses for practicing those methods. Some embodiments can also be embodied in the form of program code recorded in tangible media, such as magnetic recording media, optical recording media, solid state memory, floppy diskettes, CD-ROMs, hard drives, or any other non-transitory machine-readable storage medium, wherein, when the program code is loaded into and executed by a machine, such as a computer, the machine becomes an apparatus for practicing the patented invention(s). Some embodiments can also be embodied in the form of program code, for example, stored in a non-transitory machine-readable storage medium including being loaded into and/or executed by a machine, wherein, when the program code is loaded into and executed by a machine, such as a
15 computer or a processor, the machine becomes an apparatus for practicing the patented invention(s). When implemented on a general-purpose processor, the program code segments combine with the processor to provide a unique device that operates analogously to specific logic circuits.

[00148] Unless explicitly stated otherwise, each numerical value and range should be interpreted as being approximate as if the word “about” or “approximately” preceded the value or range.

20 [00149] The use of figure numbers and/or figure reference labels in the claims is intended to identify one or more possible embodiments of the claimed subject matter in order to facilitate the interpretation of the claims. Such use is not to be construed as necessarily limiting the scope of those claims to the embodiments shown in the corresponding figures.

25 [00150] Although the elements in the following method claims, if any, are recited in a particular sequence with corresponding labeling, unless the claim recitations otherwise imply a particular sequence for implementing some or all of those elements, those elements are not necessarily intended to be limited to being implemented in that particular sequence.

[00151] Reference herein to “one embodiment” or “an embodiment” means that a particular feature, structure, or characteristic described in connection with the embodiment can be included in

at least one embodiment of the disclosure. The appearances of the phrase “in one embodiment” in various places in the specification are not necessarily all referring to the same embodiment, nor are separate or alternative embodiments necessarily mutually exclusive of other embodiments. The same applies to the term “implementation.”

5 **[00152]** Unless otherwise specified herein, the use of the ordinal adjectives “first,” “second,” “third,” etc., to refer to an object of a plurality of like objects merely indicates that different instances of such like objects are being referred to, and is not intended to imply that the like objects so referred-to have to be in a corresponding order or sequence, either temporally, spatially, in ranking, or in any other manner.

10 **[00153]** Unless otherwise specified herein, in addition to its plain meaning, the conjunction “if” may also or alternatively be construed to mean “when” or “upon” or “in response to determining” or “in response to detecting,” which construal may depend on the corresponding specific context. For example, the phrase “if it is determined” or “if [a stated condition] is detected” may be construed to mean “upon determining” or “in response to determining” or “upon detecting [the stated condition
15 or event]” or “in response to detecting [the stated condition or event].”

[00154] Also for purposes of this description, the terms “couple,” “coupling,” “coupled,” “connect,” “connecting,” or “connected” refer to any manner known in the art or later developed in which energy is allowed to be transferred between two or more elements, and the interposition of one or more additional elements is contemplated, although not required. Conversely, the terms
20 “directly coupled,” “directly connected,” etc., imply the absence of such additional elements.

[00155] As used herein in reference to an element and a standard, the term compatible means that the element communicates with other elements in a manner wholly or partially specified by the standard and would be recognized by other elements as sufficiently capable of communicating with the other elements in the manner specified by the standard. The compatible element does not need
25 to operate internally in a manner specified by the standard.

[00156] The functions of the various elements shown in the figures, including any functional blocks labeled as “processors” and/or “controllers,” may be provided through the use of dedicated hardware as well as hardware capable of executing software in association with appropriate software. When provided by a processor, the functions may be provided by a single dedicated

processor, by a single shared processor, or by a plurality of individual processors, some of which may be shared. Moreover, explicit use of the term “processor” or “controller” should not be construed to refer exclusively to hardware capable of executing software, and may implicitly include, without limitation, digital signal processor (DSP) hardware, network processor, application specific integrated circuit (ASIC), field programmable gate array (FPGA), read only memory (ROM) for storing software, random access memory (RAM), and nonvolatile storage. Other hardware, conventional and/or custom, may also be included. Similarly, any switches shown in the figures are conceptual only. Their function may be carried out through the operation of program logic, through dedicated logic, through the interaction of program control and dedicated logic, or even manually, the particular technique being selectable by the implementer as more specifically understood from the context.

[00157] As used in this application, the terms “circuit,” “circuitry” may refer to one or more or all of the following: (a) hardware-only circuit implementations (such as implementations in only analog and/or digital circuitry); (b) combinations of hardware circuits and software, such as (as applicable): (i) a combination of analog and/or digital hardware circuit(s) with software/firmware and (ii) any portions of hardware processor(s) with software (including digital signal processor(s)), software, and memory(ies) that work together to cause an apparatus, such as a mobile phone or server, to perform various functions); and (c) hardware circuit(s) and or processor(s), such as a microprocessor(s) or a portion of a microprocessor(s), that requires software (e.g., firmware) for operation, but the software may not be present when it is not needed for operation.” This definition of circuitry applies to all uses of this term in this application, including in any claims. As a further example, as used in this application, the term circuitry also covers an implementation of merely a hardware circuit or processor (or multiple processors) or portion of a hardware circuit or processor and its (or their) accompanying software and/or firmware. The term circuitry also covers, for example and if applicable to the particular claim element, a baseband integrated circuit or processor integrated circuit for a mobile device or a similar integrated circuit in server, a cellular network device, or other computing or network device.

[00158] It should be appreciated by those of ordinary skill in the art that any block diagrams herein represent conceptual views of illustrative circuitry embodying the principles of the disclosure. Similarly, it will be appreciated that any flow charts, flow diagrams, state transition diagrams, pseudo code, and the like represent various processes which may be substantially represented in

computer readable medium and so executed by a computer or processor, whether or not such computer or processor is explicitly shown.

[00159] “BRIEF SUMMARY OF SOME SPECIFIC EMBODIMENTS” in this specification is intended to introduce some example embodiments, with additional embodiments being described in
5 “DETAILED DESCRIPTION” and/or in reference to one or more drawings. “BRIEF SUMMARY OF SOME SPECIFIC EMBODIMENTS” is not intended to identify essential elements or features of the claimed subject matter, nor is it intended to limit the scope of the claimed subject matter.

CLAIMS

What is claimed is:

1. A method of jointly representing an audio signal and a video signal using a neural field, the method comprising:

5 applying positional encoding to a time stamp coordinate (t) corresponding to a video frame of the video signal to generate a higher-dimensional embedding of the time stamp coordinate (t) corresponding to the video frame and applying positional encoding to the time stamp coordinate (t) of a segment of the audio signal associated with the video signal to generate a higher-dimensional embedding of the time stamp coordinate (t) of the segment of the audio signal; and

10 training a neural-field network to represent a sequence of video frames of the video signal and an associated plurality of segments of the audio signal, the neural-field network being configured to output a respective video frame of the sequence and a respective set of audio samples representing a respective one of the associated plurality of segments in response to an input of the higher-dimensional embedding of the time stamp coordinate (t) corresponding to the video frame and the
15 higher-dimensional embedding of the time stamp coordinate (t) of the associated segment of the audio signal, the training producing a set of fixed parameters for the neural field network, wherein the neural-field network comprises a first multilayer perceptron (MLP) configured to generate a respective joint audio-video embedding in response to the higher-dimensional embedding of the time stamp coordinate (t) corresponding to the video frame and the higher-dimensional embedding
20 of the time stamp coordinate (t) of the associated segment of the audio signal, wherein the neural-field network further comprises:

a first branch configured to generate the respective video frame in response to the respective joint audio-video embedding; and

a second branch configured to generate the respective one of the associated plurality of
25 segments of the audio signal in response to the respective joint audio-video embedding.

2. The method of claim 1, wherein applying positional encoding to a time stamp coordinate (t) comprises applying a series of trigonometric functions to the time stamp coordinate (t) for fitting the output data with high-frequency components.

3. The method of claim 1 or claim 2, wherein the training comprises determining parameters for the first MLP based on backpropagating both audio losses and video losses through the first MLP.

5 4. The method of any of claims 1-3,
wherein the first branch comprises a first convolutional neural network (CNN); and
wherein the second branch comprises a second CNN.

5. The method of claim 4,
10 wherein the first branch further comprises a second MLP coupled between the first MLP and the first CNN; and
wherein the second branch further comprises a third MLP coupled between the first MLP and the second CNN.

15 6. The method of claim 5,
wherein the second and third MLP are configured to process the joint audio-video embedding generated by the first MLP as input,
wherein the first CNN is configured to process the output generated by the second MLP as input, and
20 wherein the second CNN is configured to process the output generated by third MLP as input.

7. The method of any of claims 4 to 6,
wherein the first CNN comprises a plurality of serially connected first blocks, each of the
25 first blocks including a respective convolutional layer and a respective sub-pixel convolutional layer and having processing elements configured to use a SiLU/Swish activation function; and
wherein the second CNN comprises a plurality of serially connected second blocks, each of the second blocks including a respective convolutional layer and a respective sub-pixel convolutional layer and having processing elements configured to use a Sine activation function.

30 8. The method of claim 7, wherein a total number of the second blocks is smaller than a total number of the first blocks.

9. A method of jointly representing an audio signal and a video signal using a neural field, the method comprising:

applying positional encoding to a time stamp coordinate (t) corresponding to a video frame of the video signal to generate a higher-dimensional embedding of the time stamp coordinate (t) corresponding to the video frame and applying positional encoding to the time stamp coordinate (t) of a segment of the audio signal associated with the video signal to generate a higher-dimensional embedding of the time stamp coordinate (t) of the segment of the audio signal; and

reconstructing the video frame of the video signal and generating a corresponding set of audio samples representing the associated segment of the audio signal by applying the higher-dimensional embedding of the time stamp coordinate (t) corresponding to the video frame and the higher-dimensional embedding of the time stamp coordinate (t) of the associated segment of the audio signal to a neural-field network trained to represent a sequence of video frames of the video signal and an associated plurality of segments of the audio signal, the neural-field network being configured to output a respective video frame of the sequence and a respective set of audio samples representing a respective one of the associated plurality of segments in response to an input of the higher-dimensional embedding of the time stamp coordinate (t) corresponding to the video frame and the higher-dimensional embedding of the time stamp coordinate (t) of the associated segment of the audio signal, the training producing a set of fixed parameters for the neural field network, wherein the neural-field network comprises a first multilayer perceptron (MLP) configured to generate a respective joint audio-video embedding in response to the higher-dimensional embedding of the time stamp coordinate (t) corresponding to the video frame and the higher-dimensional embedding of the time stamp coordinate (t) of the associated segment of the audio signal, wherein the neural-field network further comprises:

a first branch configured to generate the respective video frame in response to the respective joint audio-video embedding; and
a second branch configured to generate the respective one of the associated plurality of segments of the audio signal in response to the respective joint audio-video embedding.

10. The method of claim 9, wherein applying positional encoding to a time stamp coordinate (t) comprises applying a series of trigonometric functions to the time stamp coordinate (t) for fitting the output data with high-frequency components.

11. The method of claim 9 or claim 10, wherein the first MLP has been trained based on backpropagating both audio losses and video losses through the first MLP.

12. The method of any of claims 9-11,

5 wherein the first branch comprises a first convolutional neural network (CNN); and wherein the second branch comprises a second CNN.

13. The method of claim 12,

10 wherein the first branch further comprises a second MLP coupled between the first MLP and the first CNN; and

wherein the second branch further comprises a third MLP coupled between the first MLP and the second CNN.

14. The method of claim 13,

15 wherein the second and third MLP are configured to process the joint audio-video embedding generated by the first MLP as input,

wherein the first CNN is configured to process the output generated by the second MLP as input, and

20 wherein the second CNN is configured to process the output generated by third MLP as input.

15. The method of any of claims 12 to 14,

25 wherein the first CNN comprises a plurality of serially connected first blocks, each of the first blocks including a respective convolutional layer and a respective sub-pixel convolutional layer and having processing elements configured to use a SiLU/Swish activation function; and

wherein the second CNN comprises a plurality of serially connected second blocks, each of the second blocks including a respective convolutional layer and a respective sub-pixel convolutional layer and having processing elements configured to use a Sine activation function.

30 16. The method of claim 15, wherein a total number of the second blocks is smaller than a total number of the first blocks.

17. A non-transitory computer-readable medium storing instructions that, when executed by an electronic processor, cause the electronic processor to perform operations comprising the method of any of claims 1-16.

5 18. A method of representing an audio signal using a neural field, the method comprising:
transforming the audio signal into a pair of 2D spectrograms including a 2D magnitude spectrogram and a corresponding 2D phase spectrogram, a first dimension of the 2D spectrograms representing time, a second dimension of the 2D spectrograms representing frequency; and
training a neural-field network to represent the pair of 2D spectrograms, the neural-field
10 network being configured to output a magnitude value and a phase value corresponding to the audio signal in response to an input of a 2D coordinate, the training producing a set of fixed parameters for the neural field network.

15 19. The method of claim 18, wherein the neural-field network is implemented using a multilayer perceptron (MLP).

20 20. The method of claim 19, wherein processing elements of the MLP are configured to use an activation function selected from the group consisting of a ReLU activation function, a SiLU/Swish activation function, and a Sine activation function.

21. The method of any of claims 18 to 20, wherein the transforming comprises:
dividing the audio signal into a plurality of segments, each of the segments corresponding to a different respective time interval;
applying a Fourier transform to each of the segments to generate a corresponding spectrum;
25 splitting the corresponding spectrum into a respective magnitude component and a respective phase component;
stacking a plurality of the respective magnitude components along the first dimension to form the 2D magnitude spectrogram; and
stacking a plurality of the respective phase components along the first dimension to form the
30 2D phase spectrogram.

22. The method of claim 21, wherein the dividing comprises applying a window function to the audio signal.

23. The method of claim 21 or claim 22, wherein the plurality of segments includes overlapping
5 segments.

24. A non-transitory computer-readable medium storing instructions that, when executed by an electronic processor, cause the electronic processor to perform operations comprising the method of any of claims 18-23.

25. A method of representing an audio signal using a neural field, the method comprising:
generating a pair of 2D spectrograms including a 2D magnitude spectrogram and a
corresponding 2D phase spectrogram by testing a neural-field network that has been trained to
output a magnitude value and a phase value corresponding to the audio signal in response to an input
15 of a 2D coordinate, a first dimension of the 2D spectrograms representing time, a second dimension
of the 2D spectrograms representing frequency; and
reconstructing the audio signal based on the pair of 2D spectrograms.

26. The method of claim 25, wherein the neural-field network is implemented using a multilayer
20 perceptron (MLP).

27. The method of claim 26, wherein processing elements of the MLP are configured to use an activation function selected from the group consisting of a ReLU activation function, a SiLU/Swish activation function, and a Sine activation function.

28. The method of any of claims 25-27, wherein the reconstructing comprises:
obtaining a plurality of spectra by pairwise combining magnitude components obtained from
the 2D magnitude spectrogram and corresponding phase components obtained from the
corresponding 2D phase spectrogram;
30 applying an inverse Fourier transform to each of the spectra to generate a corresponding
time-domain segment; and

assembling a plurality of the corresponding time-segments into the reconstructed audio signal.

29. A non-transitory computer-readable medium storing instructions that, when executed by an electronic processor, cause the electronic processor to perform operations comprising the method of any of claims 25-28.

30. A method of representing an audio signal using a neural field, the method comprising:
applying positional encoding to a time stamp corresponding to a segment of the audio signal
to generate a respective higher-dimensional embedding; and

training a neural-field network to represent a plurality of segments of the audio signal, the neural-field network being configured to output a respective set of audio samples representing a respective one of the plurality of segments of the audio signal in response to an input of the respective higher-dimensional embedding, the training producing a set of fixed parameters for the neural field network.

31. The method of claim 30, wherein the neural-field network includes a multilayer perceptron (MLP).

32. The method of claim 31, wherein processing elements of the MLP are configured to use an activation function selected from the group consisting of a ReLU activation function, a SiLU/Swish activation function, and a Sine activation function.

33. The method of claim 31 or claim 32, wherein the MLP is serially connected with a convolutional neural network (CNN).

34. The method of claim 33, wherein the CNN comprises a plurality of serially connected blocks, each of the blocks including a respective convolutional layer and a respective sub-pixel convolutional layer and having processing elements configured to use a Sine activation function.

35. A non-transitory computer-readable medium storing instructions that, when executed by an electronic processor, cause the electronic processor to perform operations comprising the method of any of claims 30-34.

5 36. A method of representing an audio signal using a neural field, the method comprising:
applying positional encoding to a time stamp corresponding to a segment of the audio signal
to generate a higher-dimensional embedding; and

generating a set of audio samples representing the segment of the audio signal by applying
the higher-dimensional embedding to a neural-field network trained to represent a plurality of
10 segments of the audio signal, the neural-field network being configured to output a respective set of
audio samples representing a respective one of the plurality of segments of the audio signal in
response to an input of a respective higher-dimensional embedding.

15 37. The method of claim 36, wherein the neural-field network includes a multilayer perceptron
(MLP).

38. The method of claim 37, wherein processing elements of the MLP are configured to use an
activation function selected from the group consisting of a ReLU activation function, a SiLU/Swish
activation function, and a Sine activation function.

20 39. The method of claim 37 or claim 38, wherein the MLP is serially connected with a
convolutional neural network (CNN).

40. The method of claim 39, wherein the CNN comprises a plurality of serially connected
25 blocks, each of the blocks including a respective convolutional layer and a respective sub-pixel
convolutional layer and having processing elements configured to use a Sine activation function.

41 A non-transitory computer-readable medium storing instructions that, when executed by an
electronic processor, cause the electronic processor to perform operations comprising the method of
30 any of claims 36-40.

42. A method of jointly representing a first multimedia signal and a second multimedia signal using a neural field, the method comprising:

applying positional encoding to a time stamp corresponding to a portion of the second multimedia signal and an associated segment of the first multimedia signal to generate a respective higher-dimensional embedding; and

training a neural-field network to represent a sequence of portions of the second multimedia signal and an associated plurality of segments of the first multimedia signal, the neural-field network being configured to output a respective portion of the sequence and a respective set of samples representing a respective one of the associated plurality of segments in response to an input of the respective higher-dimensional embedding, the training producing a set of fixed parameters for the neural field network.

43. The method of claim 42, wherein the neural-field network comprises a first multilayer perceptron (MLP) configured to generate a respective joint multimedia embedding in response to the respective higher-dimensional embedding.

44. The method of claim 43, wherein the neural-field network further comprises:
a first branch configured to generate the respective portion of the sequence in response to the respective joint multimedia embedding; and

a second branch configured to generate the respective one of the associated plurality of segments in response to the respective joint multimedia embedding.

45. The method of claim 44,
wherein the first branch comprises a first convolutional neural network (CNN); and
wherein the second branch comprises a second CNN.

46. The method of claim 45,
wherein the first branch further comprises a second MLP coupled between the first MLP and the first CNN; and

wherein the second branch further comprises a third MLP coupled between the first MLP and the second CNN.

47. The method of any of claims 42-46,
wherein the first multimedia signal is a 1D signal having a first sampling rate; and
wherein the second multimedia signal is a 3D signal having a different sampling rate.

5 48. The method of claim 47,
wherein the 3D signal includes a respective 1D signal and a respective 2D signal;
wherein the respective 2D signal has a second sampling rate that is lower than the first
sampling rate; and
wherein the respective 1D signal has a third sampling rate that is lower than the second
10 sampling rate.

49. A non-transitory computer-readable medium storing instructions that, when executed by an
electronic processor, cause the electronic processor to perform operations comprising the method of
any of claims 42-48.

15

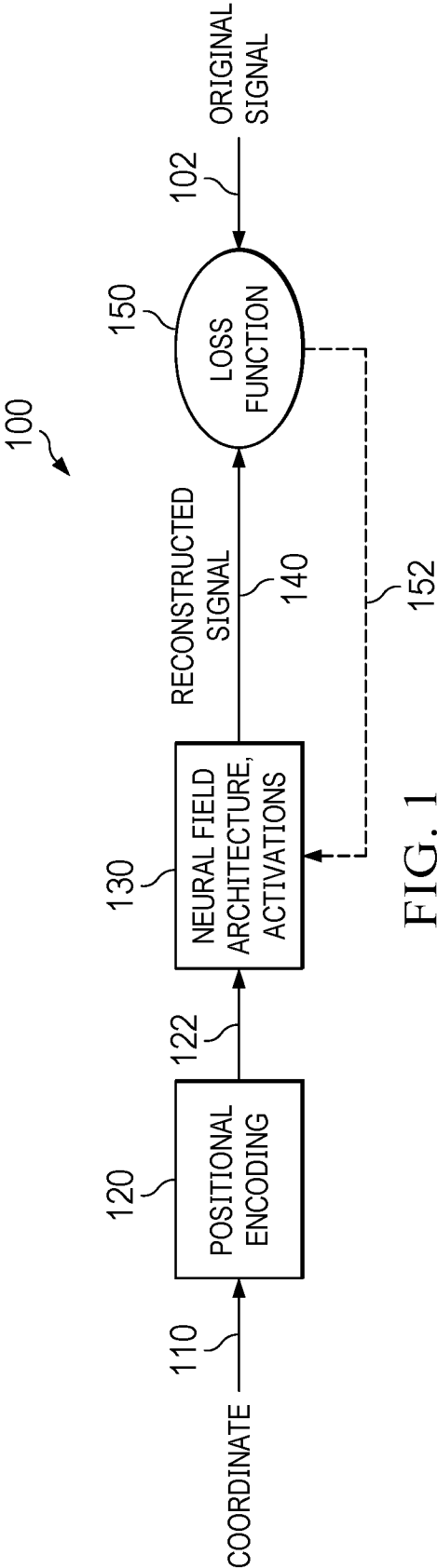


FIG. 1

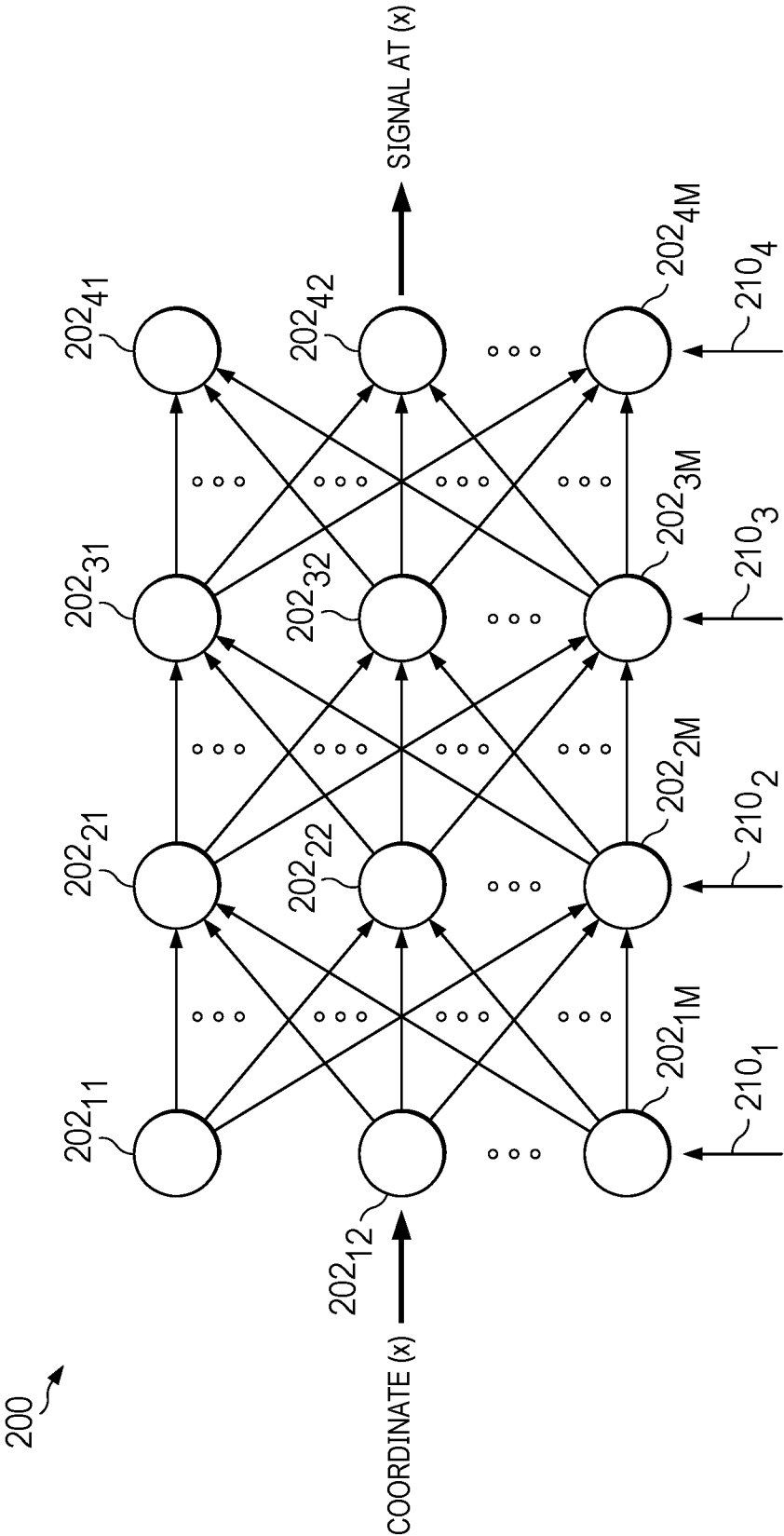


FIG. 2

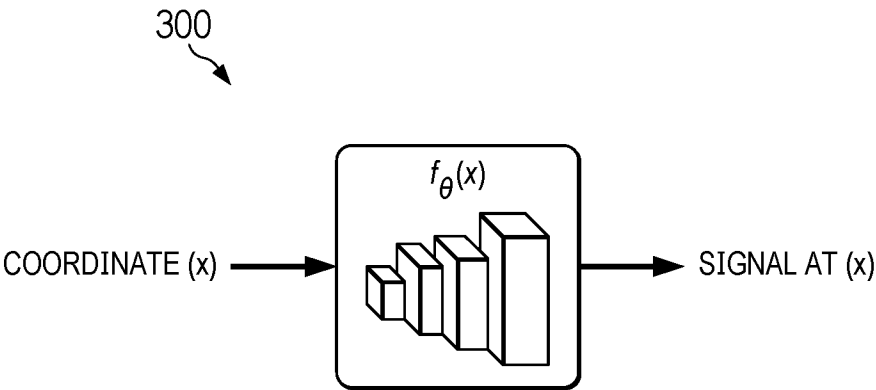


FIG. 3

4/21

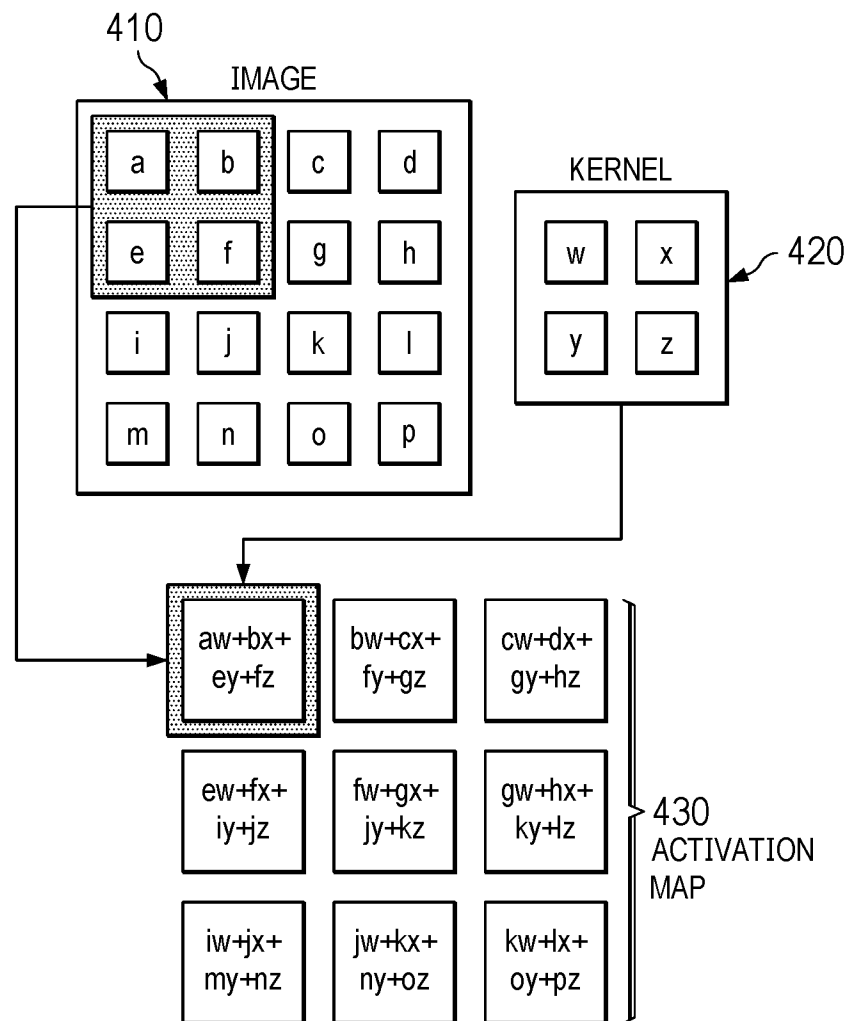


FIG. 4

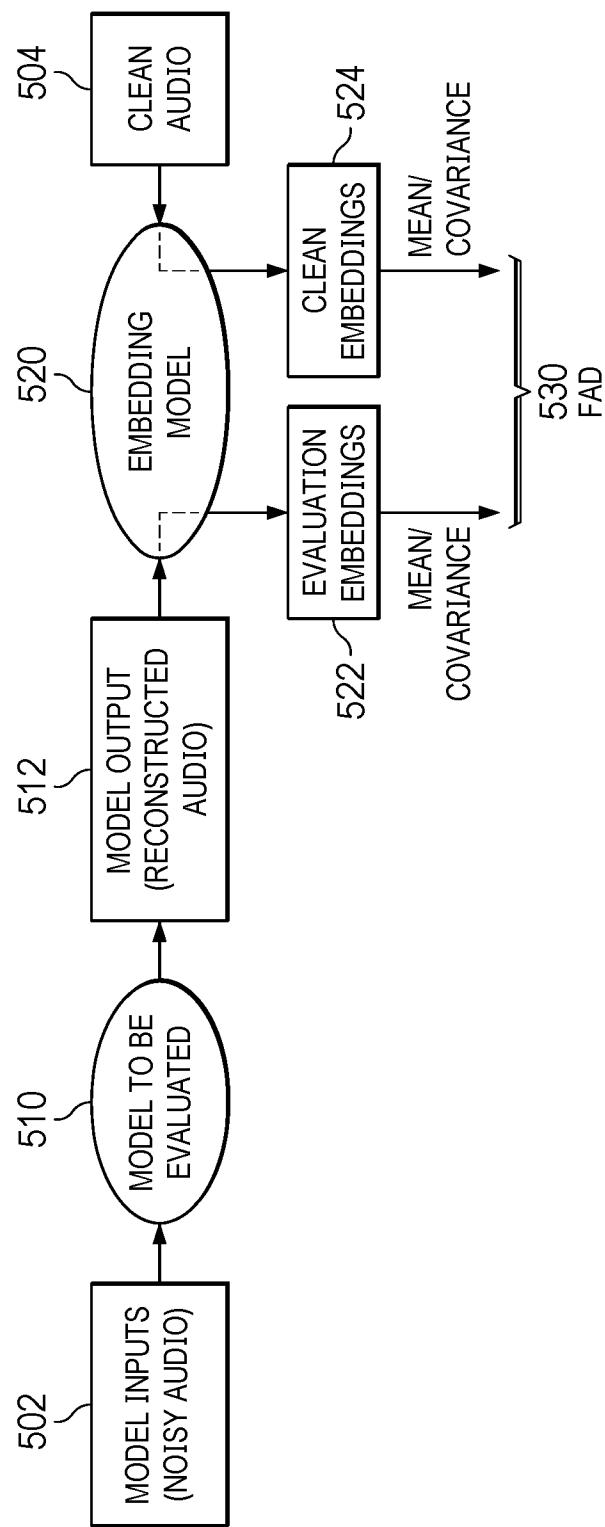


FIG. 5

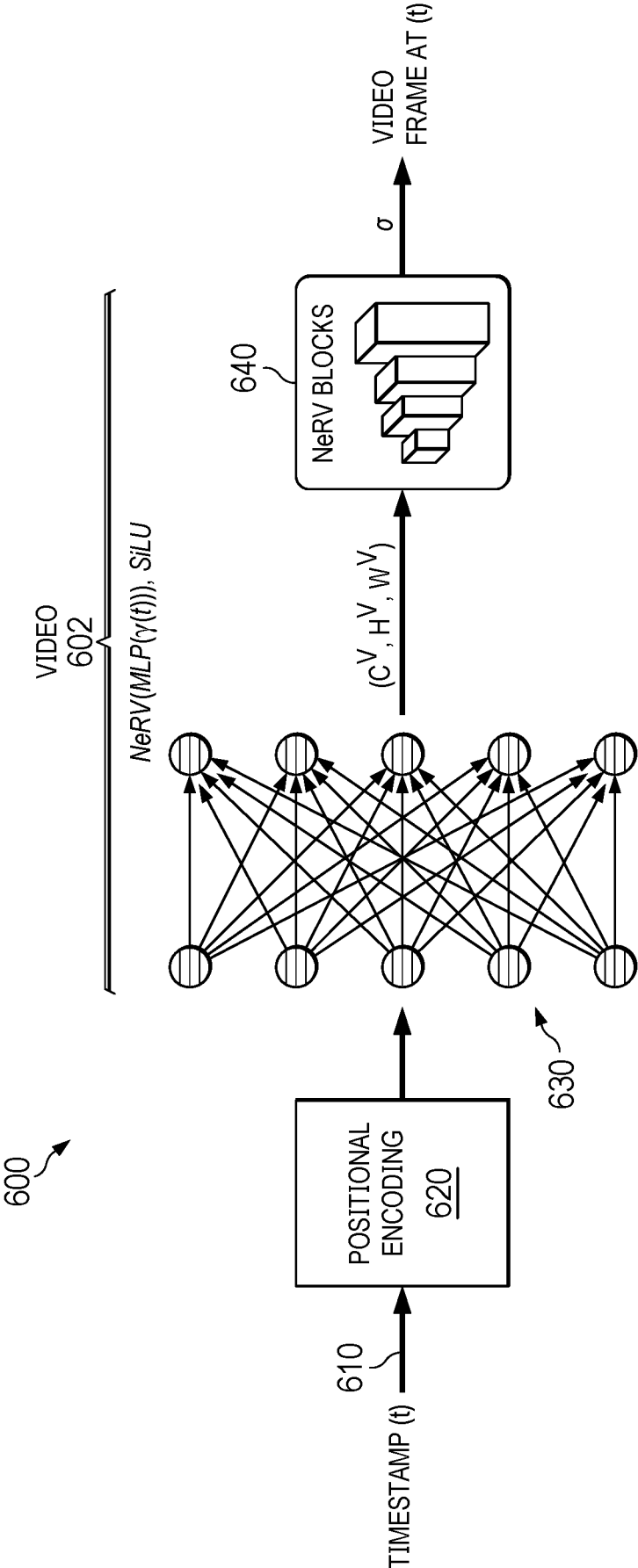


FIG. 6
(PRIOR ART)

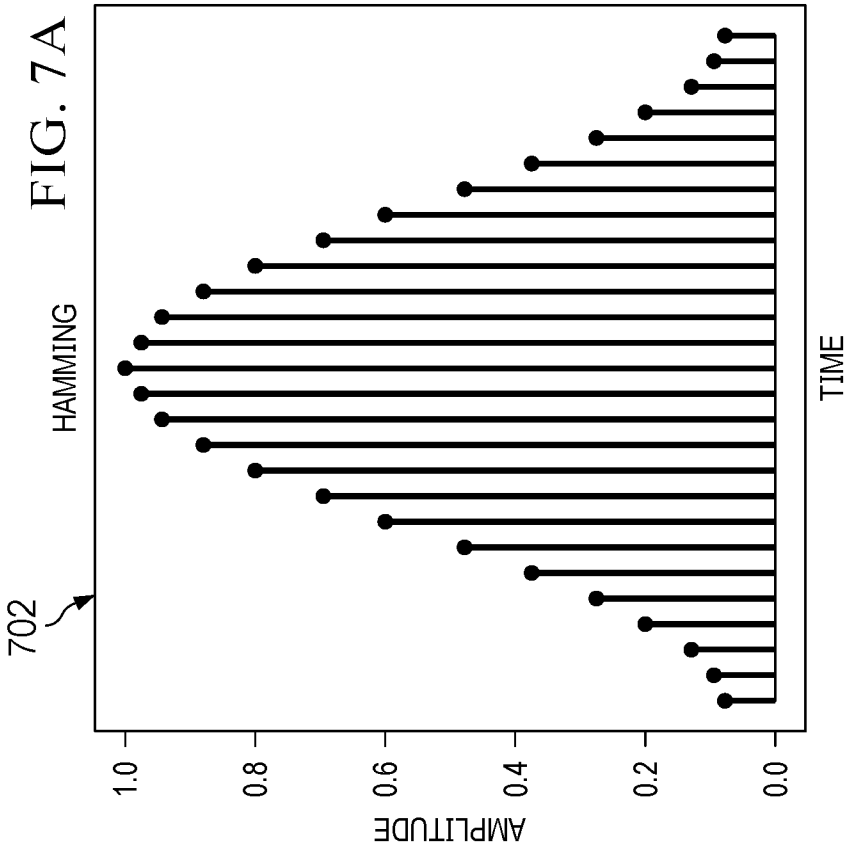
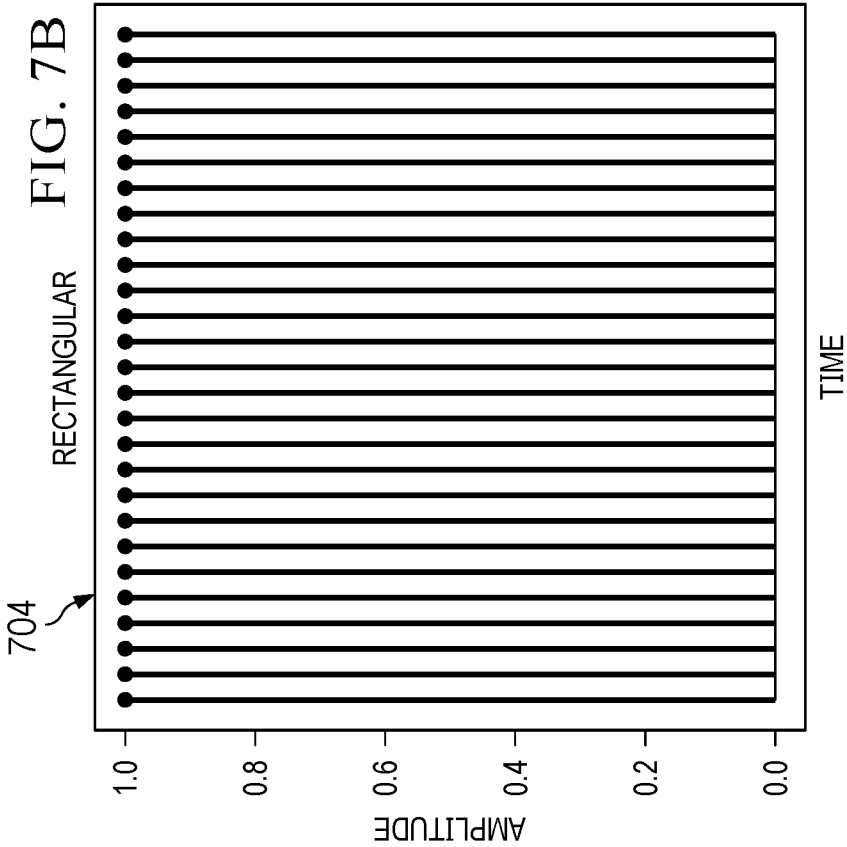
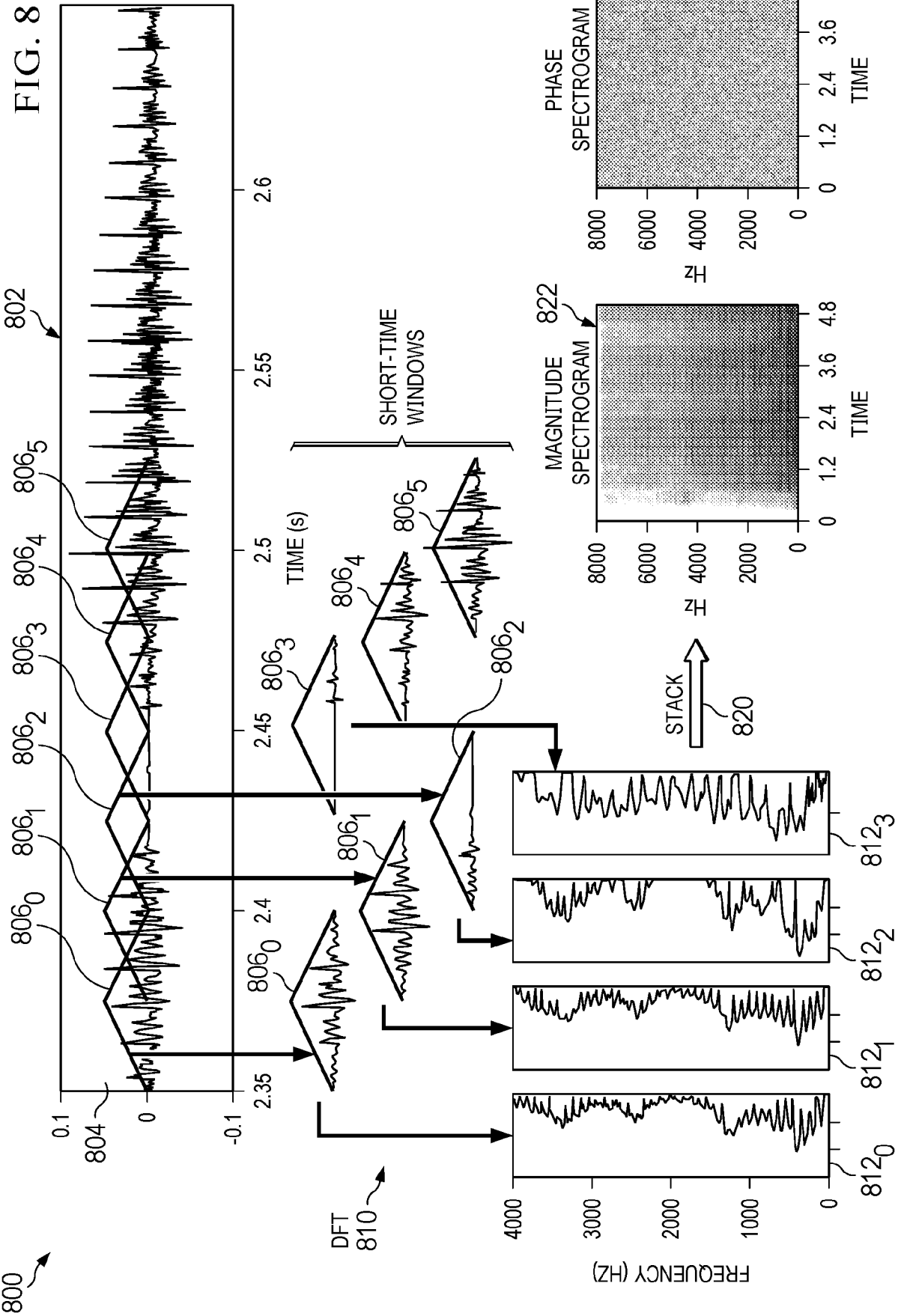


FIG. 8



9/21

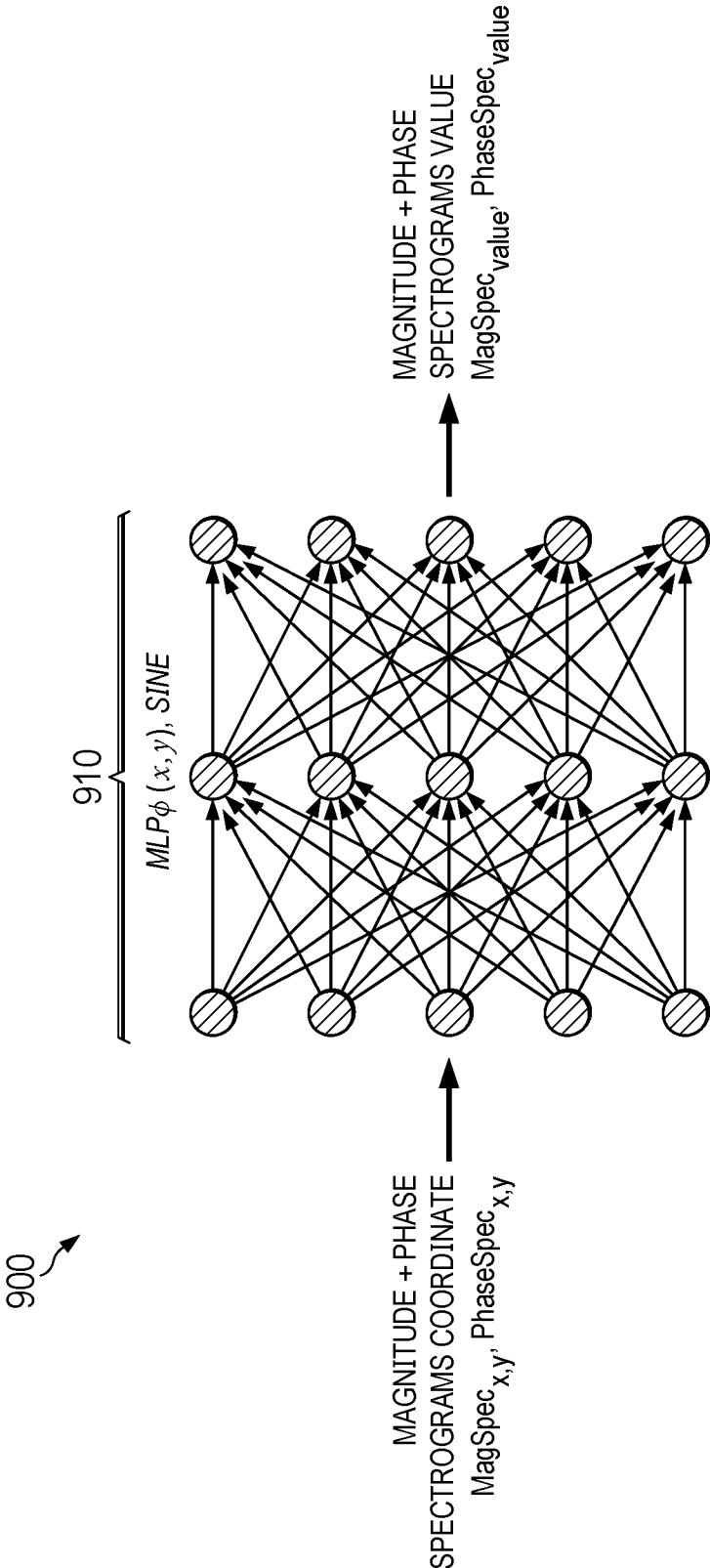


FIG. 9

10/21

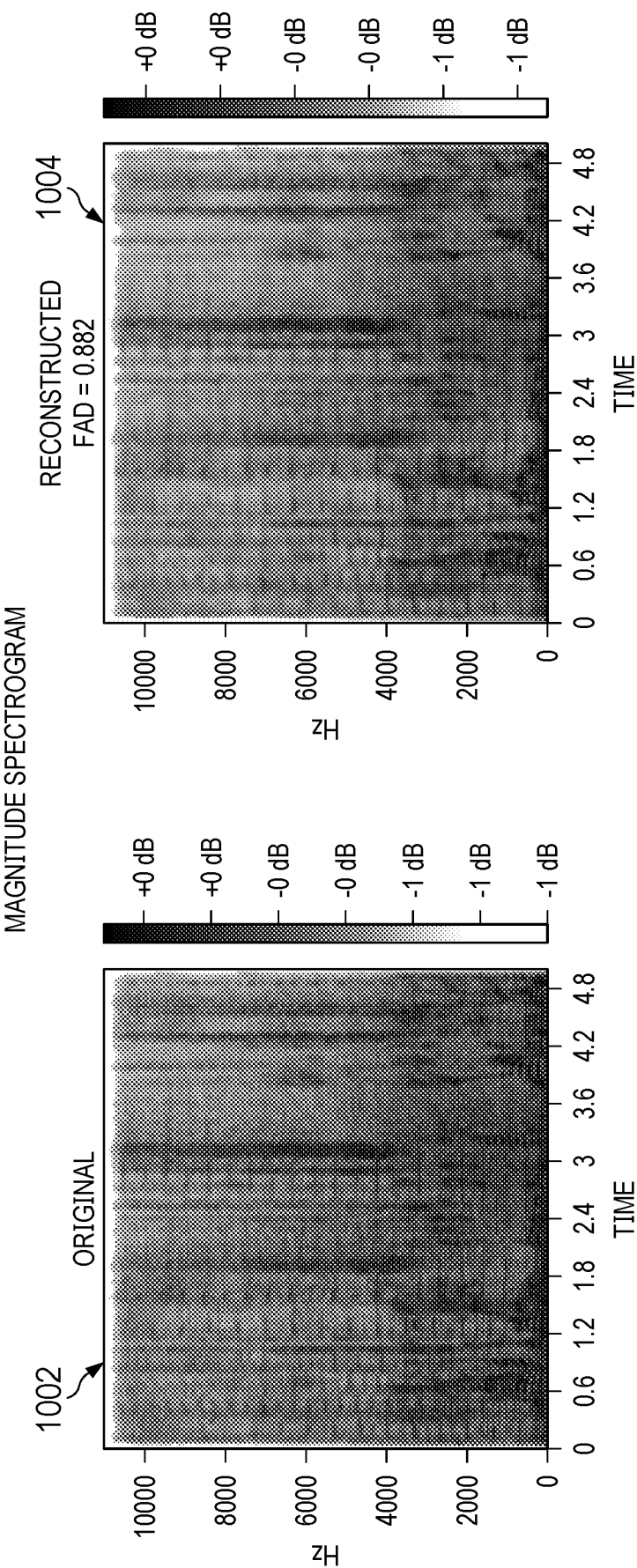


FIG. 10A

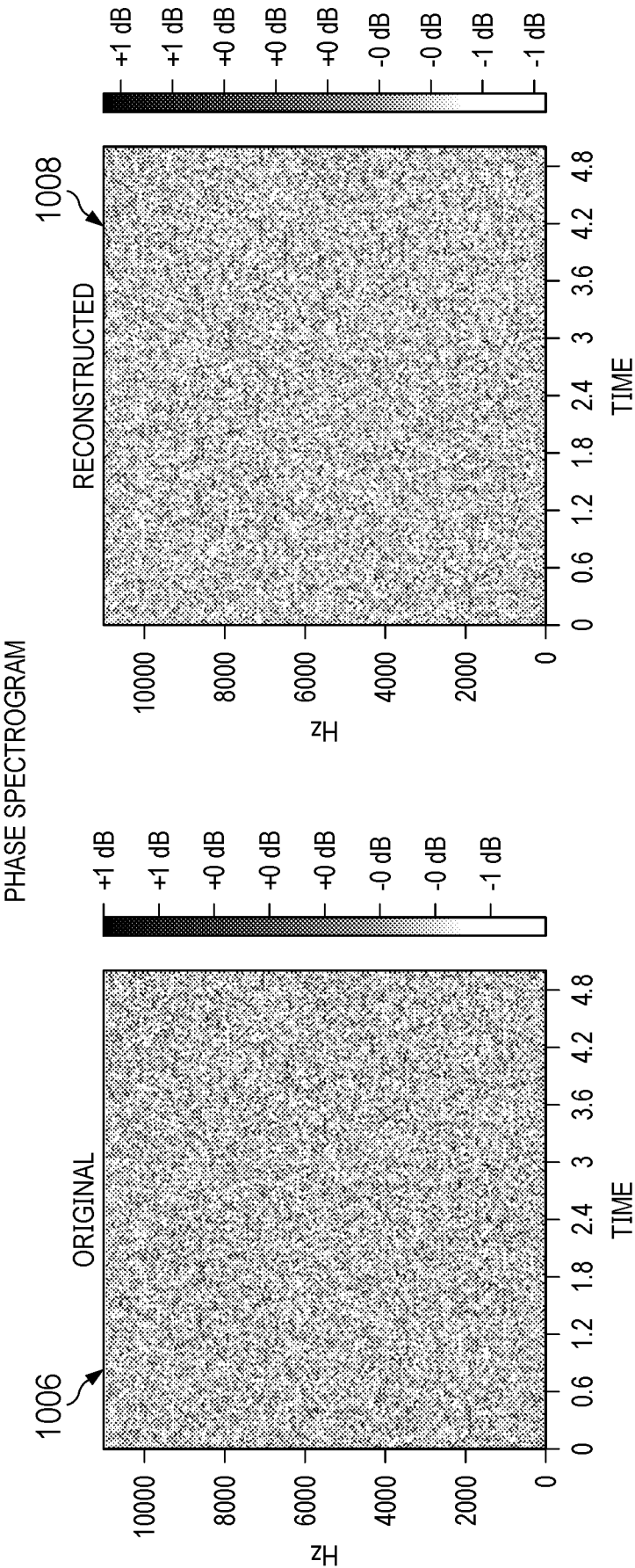


FIG. 10B

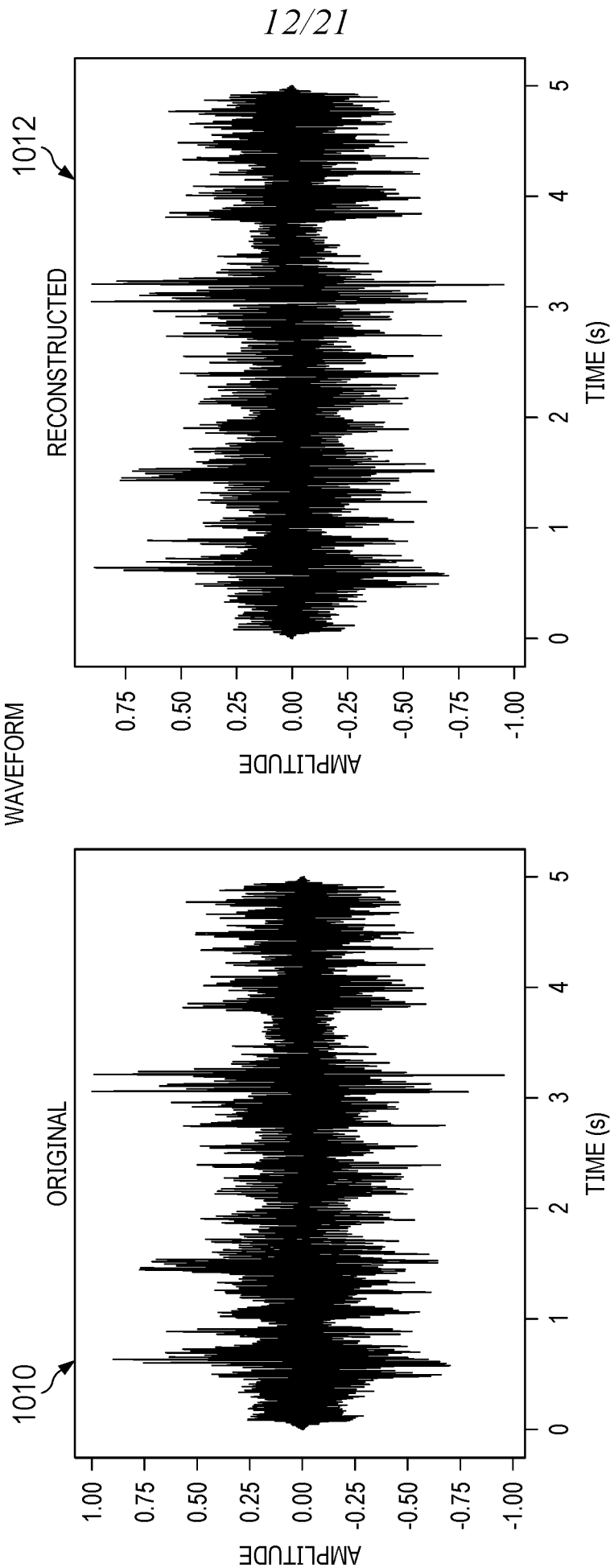


FIG. 10C

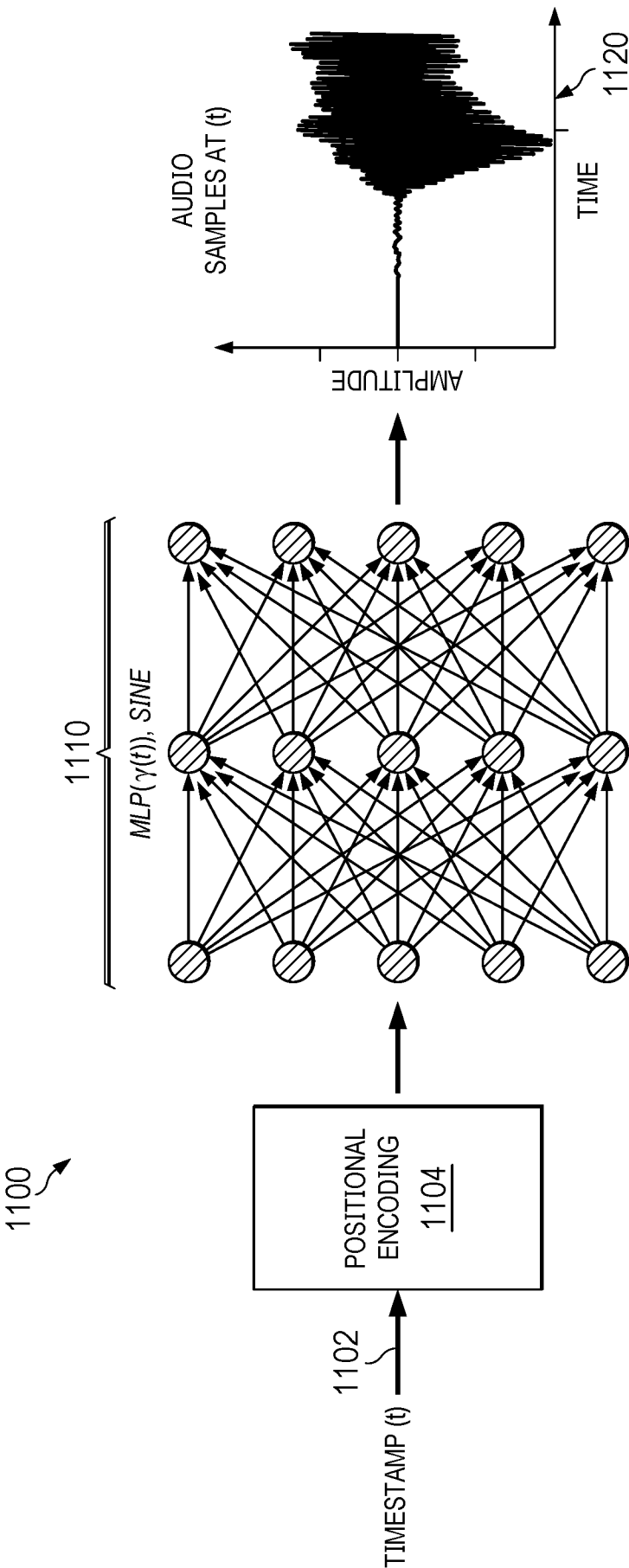
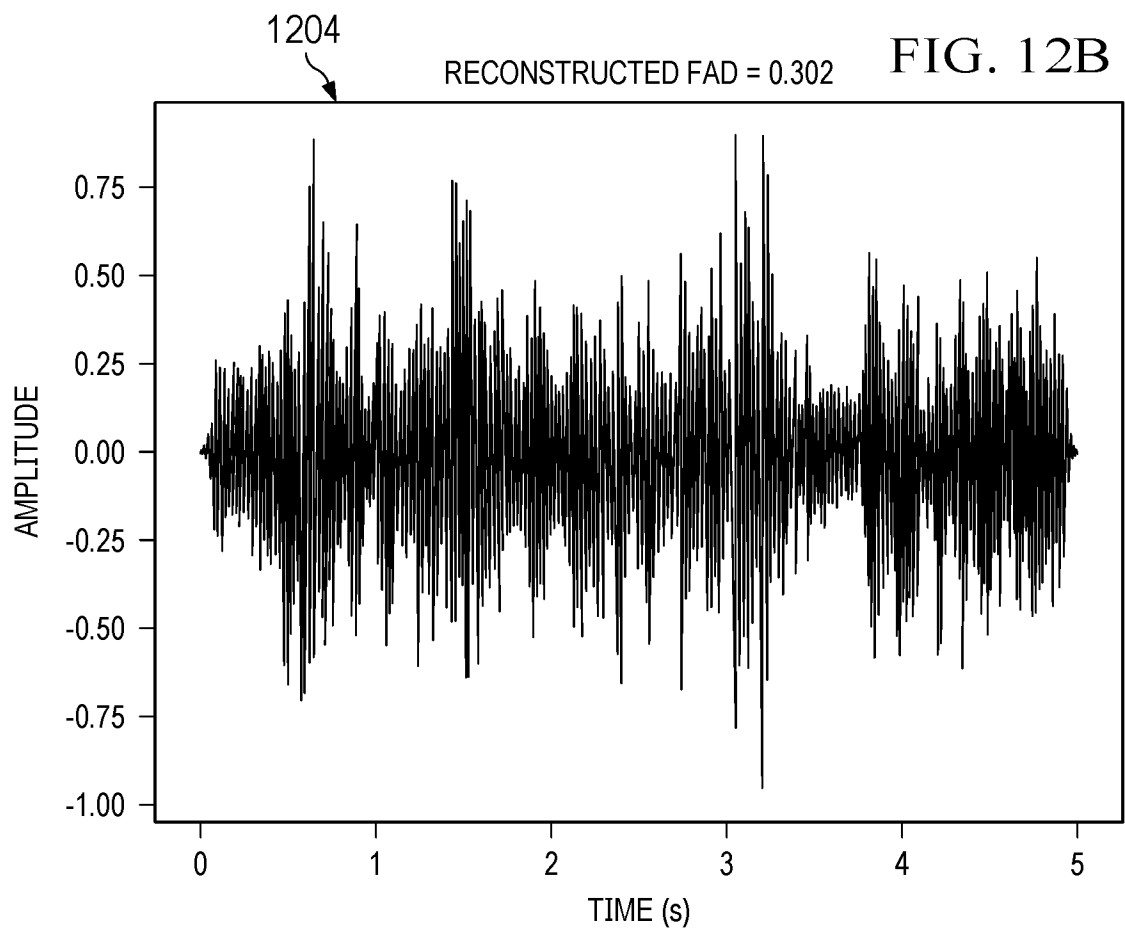
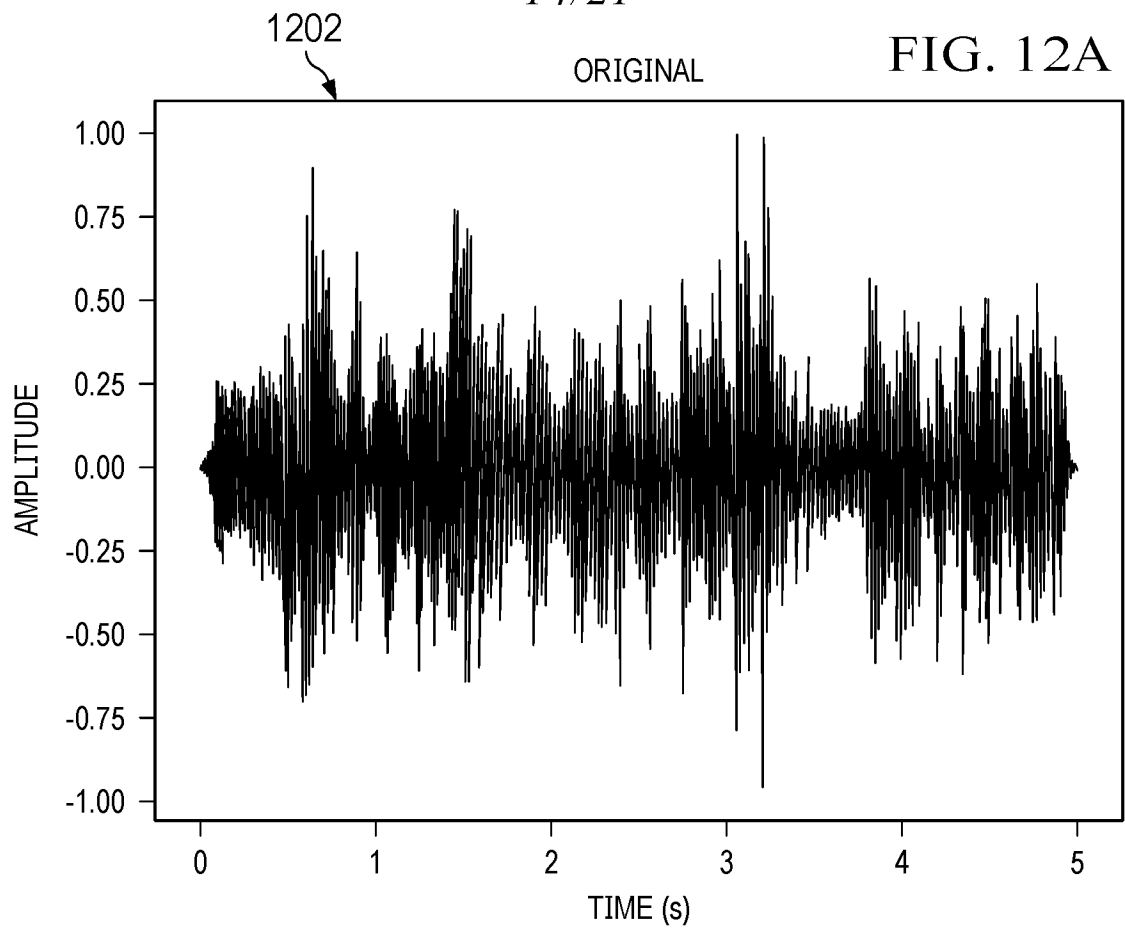


FIG. 11

14/21



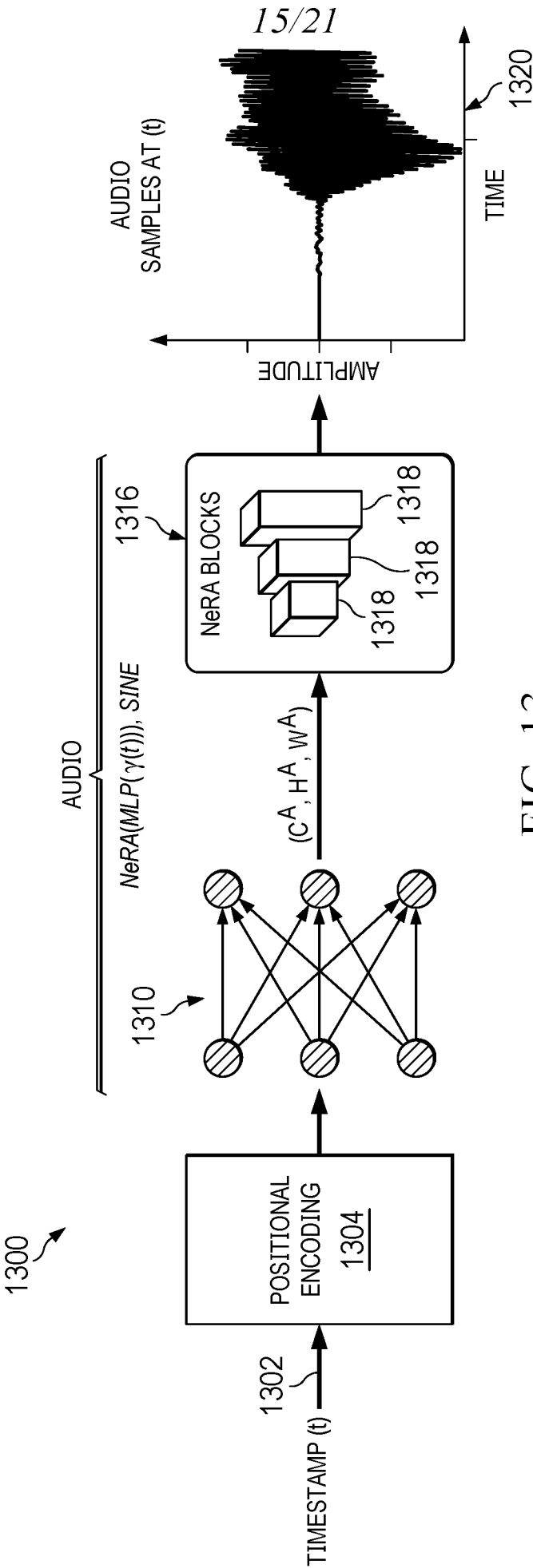


FIG. 13

16/21

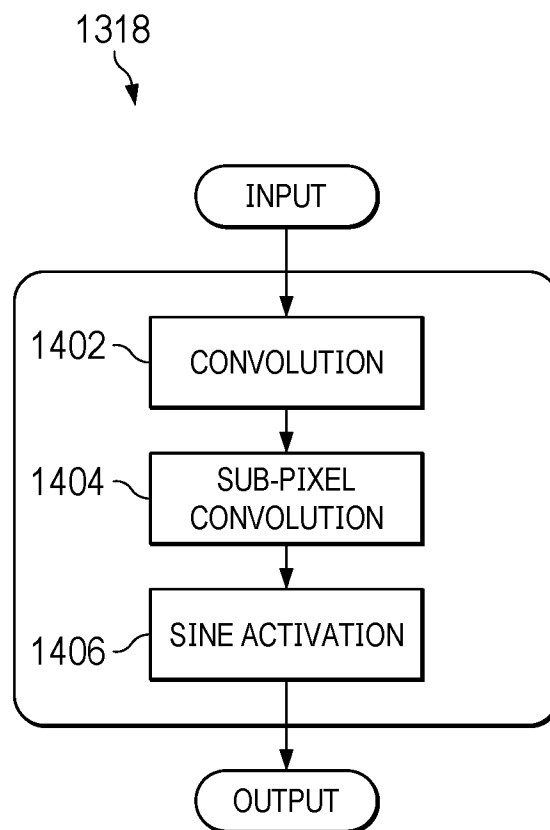
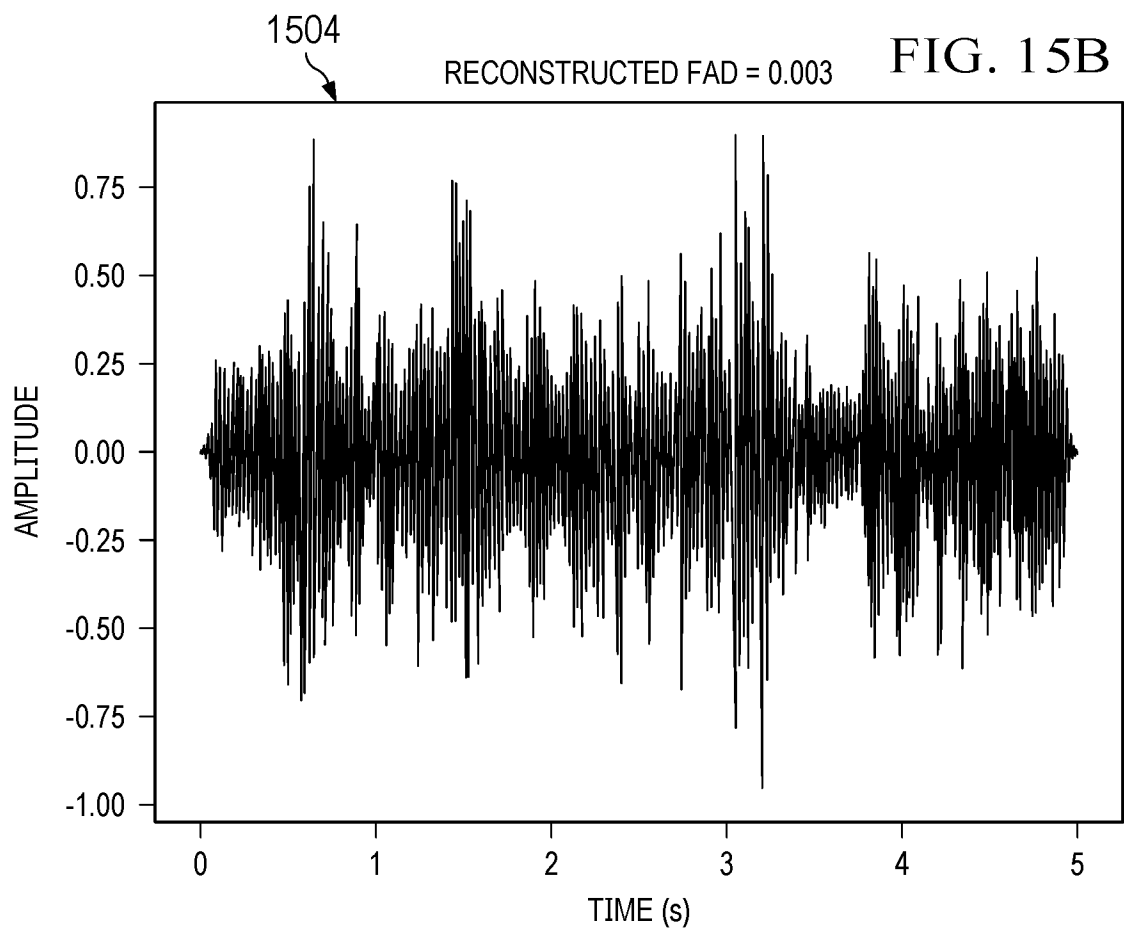
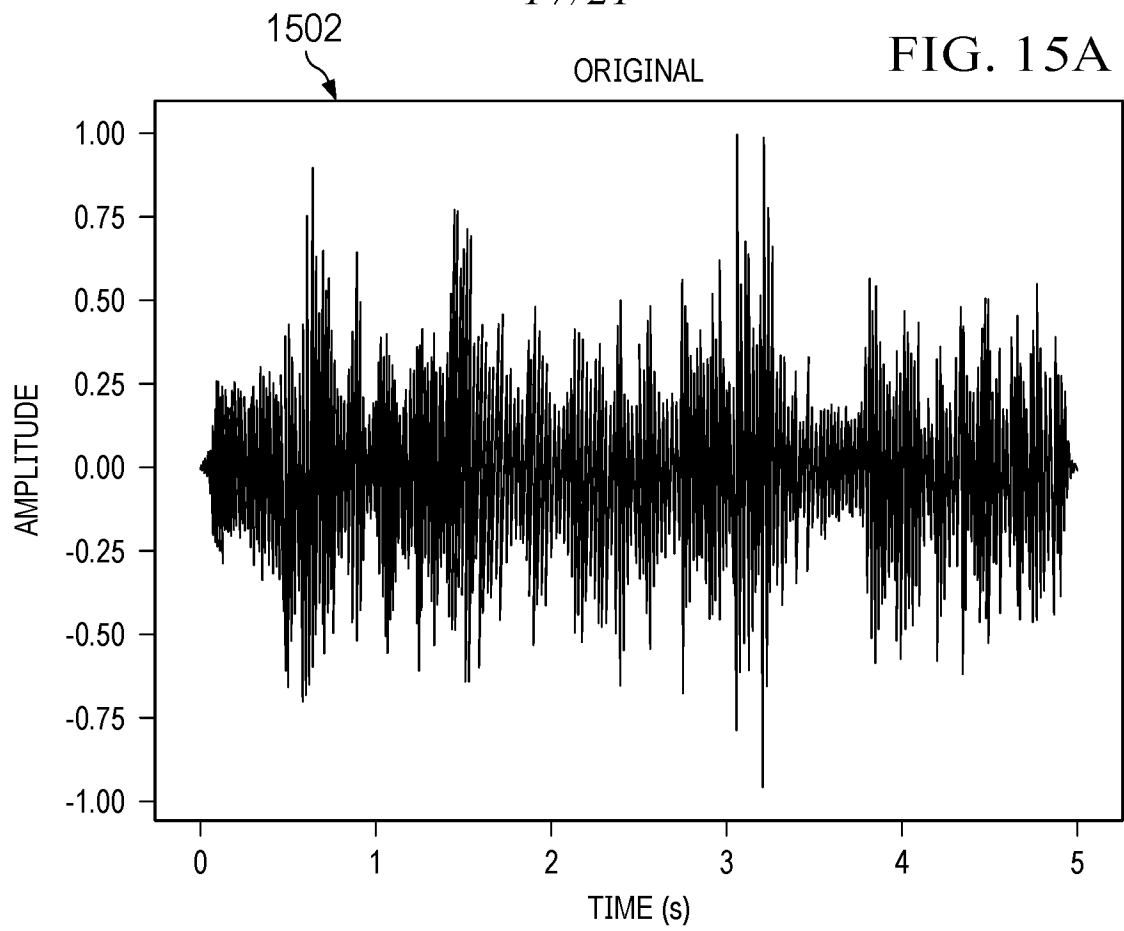


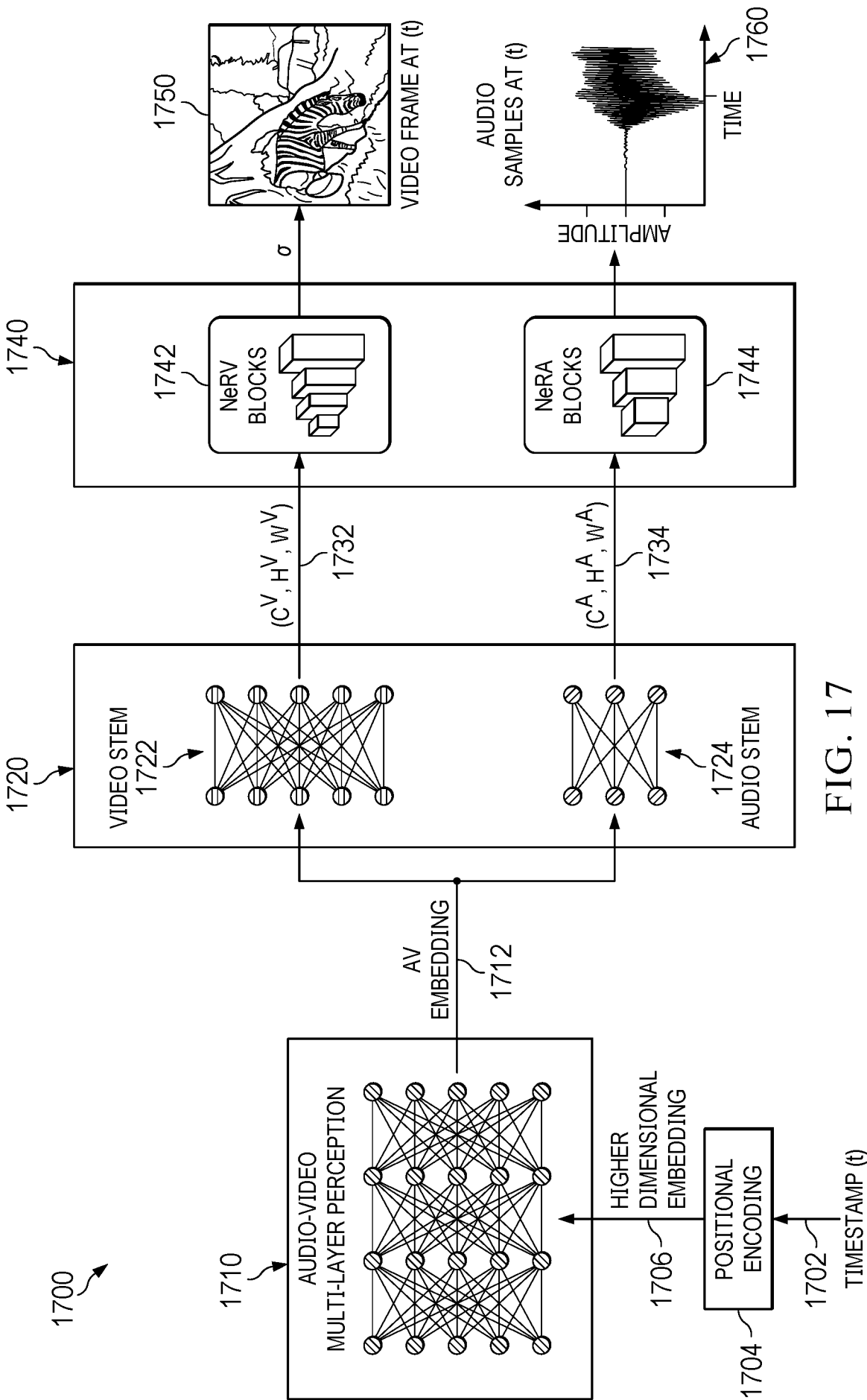
FIG. 14

17/21



AUDIO SAMPLE	FAD (MLP 1100)	FAD (NeRA 1300)
RAIN	0.028	0.0001
SPEECH	0.605	0.0845
SPEECH + MUSIC	0.302	0.003
MUSIC	0.261	0.0125

FIG. 16



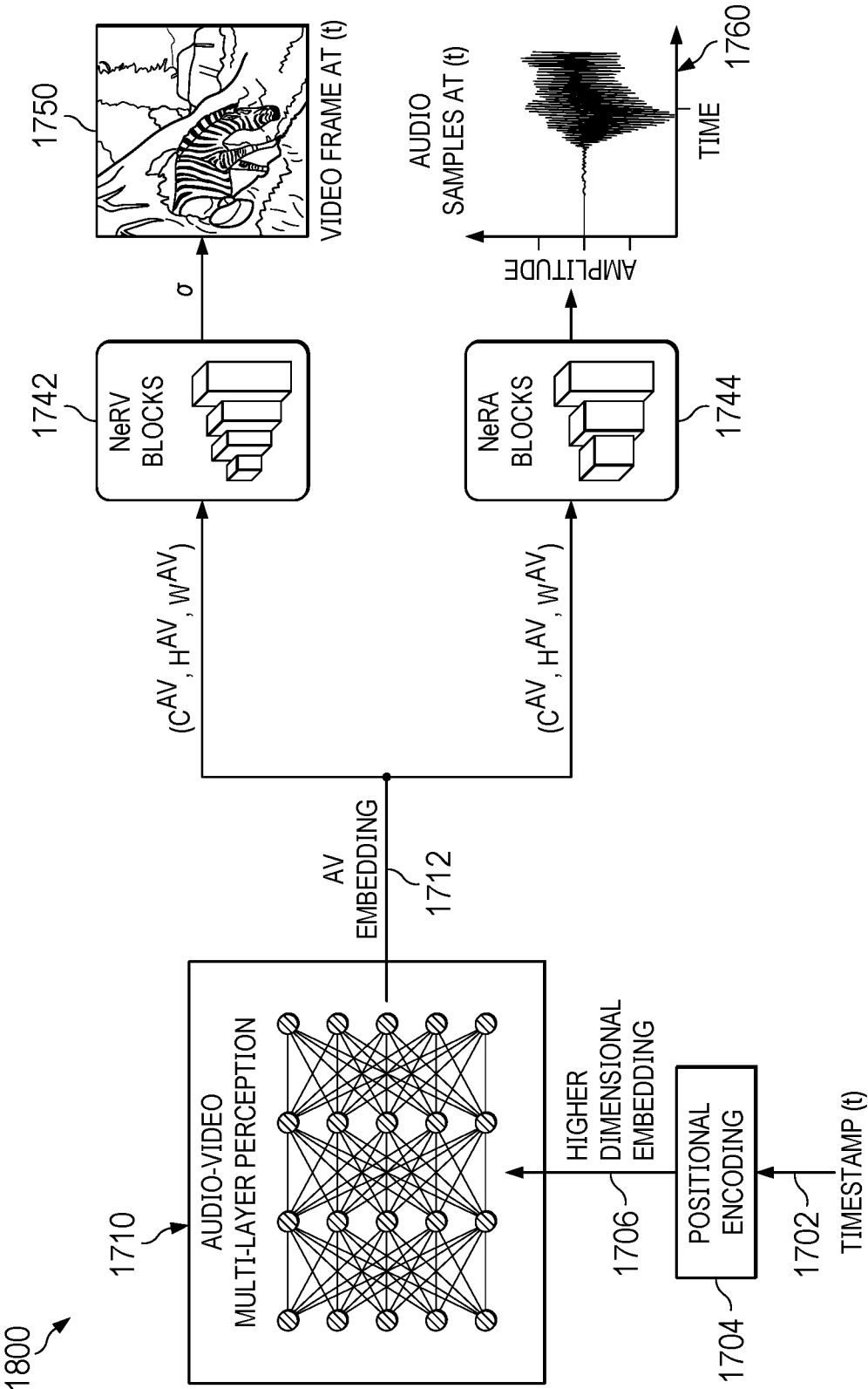


FIG. 18

21/21

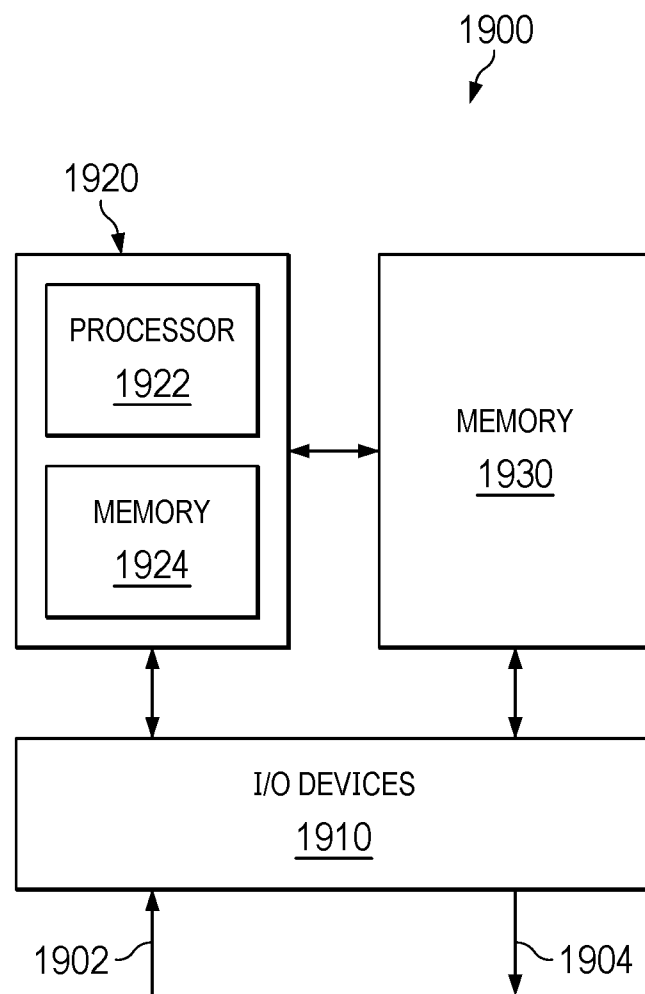


FIG. 19

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2024/054425

A. CLASSIFICATION OF SUBJECT MATTER		
INV.	H04N19/46	H04N19/68
		H04N19/85
		H04N21/8547
	G06N3/0464	G06N3/045
ADD.		
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)		
H04N G06N		
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched		
Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)		
EPO-Internal		
C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	PO-YAO HUANG ET AL: "MAViL: Masked Audio-Video Learners", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 17 July 2023 (2023-07-17), XP091563802, sections 3.1-3.3, 4.2; figures 1, 2 ----- -/-	1-17, 30-49
☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.		
* Special categories of cited documents : "A" document defining the general state of the art which is not considered to be of particular relevance "E" earlier application or patent but published on or after the international filing date "L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) "O" document referring to an oral disclosure, use, exhibition or other means "P" document published prior to the international filing date but later than the priority date claimed "T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention "X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone "Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art "&" document member of the same patent family		
Date of the actual completion of the international search		Date of mailing of the international search report
22 January 2025		25/03/2025
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016		Authorized officer Montoneri, Fabio

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2024/054425

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	GU PENGFEI ET AL: "NeRVI: Compressive neural representation of visualization images for communicating volume visualization results", COMPUTERS AND GRAPHICS, ELSEVIER, GB, vol. 116, 22 August 2023 (2023-08-22), pages 216-227, XP087437505, ISSN: 0097-8493, DOI: 10.1016/J.CAG.2023.08.024 [retrieved on 2023-08-22] section 3.1; figure 1 -----	1-17, 30-49
A	BRUNO KORBAR ET AL: "Cooperative Learning of Audio and Video Models from Self-Supervised Synchronization", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 30 June 2018 (2018-06-30), XP081426335, sections 2.1, 2.2, 2.5; figure 2 -----	1-17, 30-49
A	HARZIG PHILIPP ET AL: "Synchronized Audio-Visual Frames with Fractional Positional Encoding for Transformers in Video-to-Text Translation", 2022 IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING (ICIP), IEEE, 16 October 2022 (2022-10-16), pages 2041-2045, XP034293250, DOI: 10.1109/ICIP46576.2022.9897804 section 2.2; figure 1 -----	1-17, 30-49
A	GANGADIA DISHA: "Activation Functions: Experimentation and Comparison", 2021 6TH INTERNATIONAL CONFERENCE FOR CONVERGENCE IN TECHNOLOGY (I2CT), IEEE, 2 April 2021 (2021-04-02), pages 1-6, XP033912360, DOI: 10.1109/I2CT51068.2021.9417890 section III -----	1-15

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2024/054425

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims;; it is covered by claims Nos.:
1 - 17, 30 - 49

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-17, 30-49

relating to the use of time-stamp based positional embedding
for neural field representation of multimedia signals

1.1. claims: 1-17, 42-49

relating to the joint representation of audio and video
signals

1.2. claims: 30-41

relating to the representation of an audio signal

2. claims: 18-29

relating to the use of Short-Time Fourier Transform for
neural field representation of audio signals
