(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) **EP 3 680 901 A1**

(12) **EUROPEAN PATENT APPLICATION**

(43) Date of publication:
15.07.2020 Bulletin 2020/29

(21) Application number: **19150642.7**

(22) Date of filing: **08.01.2019**

(51) Int Cl.:
G10L 21/0208 *(2013.01)* ***G10L 25/30*** *(2013.01)*
G10L 21/0232 *(2013.01)* ***G10L 21/0264*** *(2013.01)*

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR
Designated Extension States:
BA ME
Designated Validation States:
KH MA MD TN

(71) Applicant: **Huawei Technologies Co., Ltd.**
Shenzhen, Guangdong 518129 (CN)

(72) Inventors:
- **KEHRENBERG, Thomas**
  94032 Passau (DE)
- **KEREN, Gil**
  94032 Passau (DE)
- **SCHULLER, Bjoern**
  94032 Passau (DE)
- **GROSCHE, Peter**
  80992 Munich (DE)
- **JIN, Wenyu**
  80992 Munich (DE)

(74) Representative: **Pfenning, Meinig & Partner mbB**
Patent- und Rechtsanwälte
Joachimsthaler Straße 10-12
10719 Berlin (DE)

(54) **A SOUND PROCESSING APPARATUS AND METHOD**

(57) The invention relates to a sound processing apparatus (100) configured to process a sound signal comprising a target signal and a current noise signal into an enhanced sound signal. The apparatus (100) comprises processing circuitry (101) configured to provide a neural network (103). The neural network (103) has been trained, in a training phase, on the basis of a plurality of training signals, including a plurality of training sound signals and a plurality of training target signals, using a first set of features based on the plurality of training signals and a second set of features based on the plurality of training signals. In an application phase, the neural network (103) is configured to process the sound signal into the enhanced sound signal using the first set of features, but not the second set of features. Moreover, the invention relates to a corresponding sound processing method.

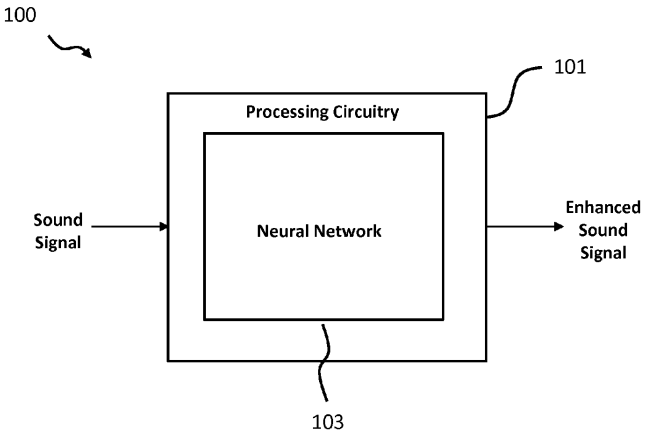


Fig. 1

EP 3 680 901 A1

Description

TECHNICAL FIELD

[0001] The present invention relates to sound processing. In particular, the present invention relates to a sound processing apparatus implementing a neural network for enhancing, in particular denoising, a sound signal, and to a corresponding method.

BACKGROUND

[0002] Audio enhancement based on a neural network usually makes use of only one feature set for enhancing, e.g. denoising a sound signal. As different feature sets emphasize different audio properties, techniques that enhance only one specific feature set might ignore important audio properties that exist in other feature sets.

[0003] In light of the above, there is a need for an improved sound processing apparatus as well as a corresponding sound processing method addressing at least some of the problems mentioned above.

SUMMARY

[0004] It is an object of the invention to provide an improved sound processing apparatus and a corresponding sound processing method.

[0005] The foregoing and other objects are achieved by the subject matter of the independent claims. Further implementation forms are apparent from the dependent claims, the description and the figures.

[0006] According to a first aspect the invention relates to a sound (in particular speech) processing apparatus configured to process a current noisy sound signal comprising a target signal and a current noise signal into an enhanced (e.g., de-noised) sound signal. The apparatus comprises processing circuitry configured to provide a neural network. In a training phase, the neural network has been trained (i.e. conditioned) on the basis of a plurality of training signals a plurality of corresponding training target signals, using a first set of features based on the plurality of training signals and a second set of features based on the plurality of training signals. The plurality of training signals includes a plurality of training sound signals. Each training sound signal comprises a training target signal and a training noise signal. In an application phase, the neural network is configured to process the sound signal into the enhanced sound signal using only the first set of features, but not the second set of features.

[0007] Thus, an improved sound processing apparatus is provided making use of at least two feature sets for generating the enhanced sound signal.

[0008] In a further possible implementation form of the first aspect, the processing circuitry is configured to transform the plurality of training signals from the time domain into the frequency domain, wherein the neural network, in the training phase, has been trained on the basis of the plurality of training signals in the frequency domain.

[0009] In a further possible implementation form of the first aspect, in the training phase, the neural network is further configured to generate for each of the plurality training sound signals an estimated training target signal and to generate the first set of features and the second set of features on the basis of the plurality of estimated training target signals and the plurality of training target signals.

[0010] In a further possible implementation form of the first aspect, in the training phase, the neural network is further configured to generate the first set of features on the basis of the plurality of estimated training target signals and the plurality of training target signals and the second set of features on the basis of the first set of features.

[0011] In a further possible implementation form of the first aspect, the first set of features comprises a plurality of spectra, in particular log mel spectra of the plurality of estimated training target signals and a plurality of spectra, in particular log mel spectra of the plurality of training target signals.

[0012] In a further possible implementation form of the first aspect, the second set of features comprises a first plurality of MFCC coefficients generated on the basis of the plurality of spectra of the plurality of estimated training target signals and a second plurality of MFCC coefficients generated on the basis of the plurality of spectra of the plurality of training target signals.

[0013] In a further possible implementation form of the first aspect, the neural network has been trained by simultaneously minimizing a first difference measure value between the plurality of spectra, in particular log mel spectra of the plurality of estimated training target signals and the plurality of spectra, in particular log mel spectra of the plurality of training target signals and a second difference measure value between the first plurality of MFCC coefficients and the second plurality of MFCC coefficients. In an implementation form, the first difference measure and/or the second difference measure can be the L1 loss function.

[0014] In a further possible implementation form of the first aspect, the processing circuitry is configured to generate for each of the plurality of training sound signals a two-dimensional data array, wherein each row of the two-dimensional data array contains a spectrum of the respective training sound signal, wherein the neural network is configured to generate the estimated training target signal for the respective training sound signal by removing noise from the two-dimensional data array.

[0015] In a further possible implementation form of the first aspect, the neural network is configured to remove the noise from the two-dimensional data array using image processing techniques for removing noise from an image.

[0016] In a further possible implementation form of the first aspect, the neural network comprises one or more

convolutional layers and deconvolutional layers and/or one or more skip connections.

[0017] According to a second aspect the invention relates to a corresponding sound processing method for processing a current noisy sound signal comprising a target signal and a current noise signal into an enhanced, i.e. de-noised sound signal. The method comprises the steps of: providing a neural network, wherein, in a training phase, the neural network has been trained on the basis of a plurality of training signals, including a plurality of training sound signals, wherein each training sound signal comprises a training target signal and a training noise signal, and on the basis of a plurality of corresponding training target signals, using a first set of features based on the plurality of training signals and a second set of features based on the plurality of training signals; and, in an application phase, processing the sound signal into the enhanced sound signal using only the first set of features, but not the second set of features.

[0018] Thus, an improved sound processing method is provided making use of at least two feature sets for generating the enhanced sound signal.

[0019] The sound processing method according to the second aspect of the invention can be performed by the sound processing apparatus according to the first aspect of the invention. Further features of the sound processing method according to the second aspect of the invention result directly from the functionality of the sound processing apparatus according to the first aspect of the invention and its different implementation forms described above and below.

[0020] According to a third aspect the invention relates to a computer program product comprising program code for performing the method according to the second aspect when executed on a computer.

[0021] Details of one or more embodiments are set forth in the accompanying drawings and the description below. Other features, objects, and advantages will be apparent from the description, drawings, and claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0022] In the following embodiments of the invention are described in more detail with reference to the attached figures and drawings, in which:

Fig. 1 is a schematic diagram showing an example of a sound processing apparatus according to an embodiment of the invention;

Fig. 2 is a schematic diagram illustrating in more detail different processing blocks and/or steps implemented in a training phase of a sound processing apparatus according to an embodiment of the invention;

Fig. 3 is a schematic diagram illustrating in more detail different processing blocks and/or steps implemented in a training phase of a sound processing apparatus according to an embodiment of the invention; and

Fig. 4 is a flow diagram showing an example of a sound processing method according to an embodiment of the invention.

[0023] In the following identical reference signs refer to identical or at least functionally equivalent features.

DETAILED DESCRIPTION OF THE EMBODIMENTS

[0024] In the following description, reference is made to the accompanying figures, which form part of the disclosure, and which show, by way of illustration, specific aspects of embodiments of the invention or specific aspects in which embodiments of the present invention may be used. It is understood that embodiments of the invention may be used in other aspects and comprise structural or logical changes not depicted in the figures. The following detailed description, therefore, is not to be taken in a limiting sense, and the scope of the present invention is defined by the appended claims.

[0025] For instance, it is to be understood that a disclosure in connection with a described method may also hold true for a corresponding device or system configured to perform the method and vice versa. For example, if one or a plurality of specific method steps are described, a corresponding device may include one or a plurality of units, e.g., functional units, to perform the described one or plurality of method steps (e.g., one unit performing the one or plurality of steps, or a plurality of units each performing one or more of the plurality of steps), even if such one or more units are not explicitly described or illustrated in the figures. On the other hand, for example, if a specific apparatus is described based on one or a plurality of units, e.g., functional units, a corresponding method may include one step to perform the functionality of the one or plurality of units (e.g., one step performing the functionality of the one or plurality of units, or a plurality of steps each performing the functionality of one or more of the plurality of units), even if such one or plurality of steps are not explicitly described or illustrated in the figures. Further, it is understood that the features of the various exemplary embodiments and/or aspects described herein may be combined with each other, unless specifically noted otherwise.

[0026] Figure 1 shows a sound processing apparatus 100 according to an embodiment of the invention. The sound processing apparatus 100 is configured to process a current noisy sound signal or waveform, e.g., a speech signal, comprising a target signal and a current noise signal into an enhanced, i.e. de-noised sound signal or waveform. The sound processing apparatus 100, which can be implemented, for instance, as a mobile phone, can comprise a microphone for recording the current noisy sound signal and a loudspeaker for reproducing

the enhanced sound signal. The current noisy sound signal and the enhanced, i.e. de-noised sound signal can be mono sound signals.

[0027] As illustrated in figure 1, the sound processing apparatus 100 comprises processing circuitry 101, configured to provide a deep neural network 103. The deep neural network 103 can comprise one or more convolutional layers and/or deconvolutional layers and/or one or more skip connections. The processing circuitry may comprise, for example, one or more processors (e.g., microprocessors) and a non-transitory storage medium carrying a program code that can be read and executed by the one or more processors. The program code when executed by the one or more processors causes the sound processing apparatus 100 to carry out the methods and operations described herein.

[0028] As will be described in more detail below, in a training phase, the neural network 103 is configured, to be trained and has been trained, i.e. conditioned on the basis of a plurality of training signals, including a plurality of training sound signals, wherein each training sound signal comprises, i.e. is a combination of a training target signal and a training noise signal, and a plurality of corresponding training target signals, using a first set of features and a second set of features based on the plurality of training signals. Once trained in such a way the neural network 103 is further configured, in an application phase, to process the sound signal into the enhanced sound signal using only the first set of features, but not the second set of features.

[0029] In an embodiment, the processing circuitry 101 is configured to transform the plurality of training signals from the time domain into the frequency domain, wherein the neural network 103 is configured, in the training phase, to be trained and has been trained on the basis of the plurality of training signals in the frequency domain.

[0030] In an embodiment, the neural network 103 is further configured, in the training phase, to generate for each of the plurality training sound signals an estimated training target signal and to generate the first set of features and the second set of features on the basis of the plurality of estimated training target signals and the plurality of training target signals.

[0031] In an embodiment, the neural network 103 is further configured, in the training phase, to generate the first set of features on the basis of the plurality of estimated training target signals and the plurality of training target signals and the second set of features on the basis of the first set of features. In other words, according to an embodiment the neural network 103 is configured to derive the second set of features from the first set of features extracted from the plurality of training signals.

[0032] Figures 2 and 3 illustrate in more detail different processing blocks and/or steps implemented in the sound processing apparatus 100 according to an embodiment in the training phase and the application phase, respectively.

[0033] In blocks 201 and 202 of figure 2 the processing circuitry 101 is configured to convert a respective training target signal (referred to as noisy waveform in figure 2) and a respective training target signal (referred to as clean waveform in figure 2) into a respective log Mel spectrum thereof. These spectra define the first set of features and can be based on temporal segments of the respective training target signal and the respective training target signal, such as temporal segments having a length of 2.5 seconds. In an embodiment, the processing circuitry 101 is configured to generate the respective log Mel spectra using a FFT/DFT with a large number of Fourier coefficients, such as 1024 Fourier coefficients, and a large number of Mel filterbanks, such as 256 Mel filterbanks.

[0034] In block 203 of figure 2 the respective log Mel spectrum of the respective training target signal is processed by an encoder portion of the deep neural network 103. In an embodiment, the encoder portion of the deep neural network 103 can comprise eight convolutional layers with 2D kernel applied to the log Mel spectrogram.

[0035] In block 205 of figure 2, a decoder portion of the deep neural network 103 processes the output of block 203. In an embodiment, the decoder portion of the deep neural network 103 comprises eight deconvolutional layers with 2D kernel applied the output of block 3.

[0036] As described above, in an embodiment, the deep neural network 103 can be implemented as an encoder-decoder network. The convolutional layers of the encoder portion compress the output into a representation of a lower temporal resolution but a higher dimension at each time step. The deconvolution layers of the decoder portion work in the opposite way - they convert this representation, i.e. the output of the convolutional layers, back to the original dimensions of the sound with higher temporal resolution. According to an embodiment, one or more skip connections can exist between convolutional layer number n and deconvolutional layer number 8-n. In addition (leaky) ReLUs (rectified linear units) can be applied after each convolutional and deconvolutional layer. According to an embodiment, the last layer of the decoder portion can have no activation function. The output of block 205 is the denoised log Mel spectrum.

[0037] In block 207 of figure 2 the processing circuitry 103 is configured to extract the second set of features in the form of MFCC coefficients from a respective denoised log Mel spectra (i.e. the first set of features) provided by the neural network 103 as the output of block 205. Likewise, in block 208 of figure 2 the processing circuitry 103 is configured to extract MFCC coefficients from the respective clean log Mel spectrum provided by block 202.

[0038] In block 206 of figure 2 the respective log Mel spectra, i.e. the first set of features, provided by blocks 205 and 202 are compared on the basis of a first difference measure. Likewise, in block 210 of figure 2 the respective MFCC coefficients, i.e. the second set of features, provided by blocks 207 and 208 are compared on the basis of a second difference measure. According to an embodiment, the L1 loss function based on the abso-

lute error can be used as the first and/or second difference measure. These difference measures are optimized in the training phase for training, i.e. conditioning the neural network 103. In an embodiment, the first and second difference measure can be combined into a combined loss function. For instance, the combined loss function can be a sum of the first difference measure, e.g., the absolute difference between the log Mel spectrum of the training sound signal and the log Mel spectrum of the training target signal, and the second difference measure, e.g., the absolute difference between the MFCC coefficients extracted from the log Mel spectrum of the training sound signal and the clean MFCC coefficients, i.e. the MFCC coefficients extracted from the log Mel spectrum of the training target signal. In an embodiment, the first difference measure and the second difference measure can be weighted differently in the combined loss function. For instance, the log Mel spectrogram loss can be weighted with a factor 100 compared to the MFCC loss. In an embodiment, the neural network 103 can be trained on the basis of an Adam optimizer on default settings.

[0039] Thus, according to an embodiment, the processing circuitry 101 is configured to generate for each of the plurality of training sound signals a two-dimensional data array, wherein each row of the two-dimensional data array comprises a spectrum of the respective training sound signal, wherein the neural network 103 is configured to generate the estimated training target signal for the respective training sound signal by removing noise from the two-dimensional data array. In an embodiment, the neural network 103 is configured to remove the noise from the two-dimensional data array using image processing techniques, as disclosed, for instance, in Long and Shelhamer, "Fully Convolutional Networks for Semantic Segmentation", 2015, which is herein fully incorporated by reference.

[0040] As will be appreciated, several of the blocks/steps shown in figure 3 concerning the application phase are equivalent to the corresponding blocks/steps shown in figure 2 concerning the training phase. More specifically, blocks 301, 303 and 305 of figure 3 are identical to blocks 201, 203 and 205 of figure 2. Therefore, reference is made to the detailed description above and in the following only the differences between figures 2 and 3 will be described in more detail.

[0041] In block 302 of figure 3 the processing circuitry 101 is configured to extract phase information from the current noisy sound signal (referred to as noisy waveform in figure 3). This noisy phase information is used in block 307 of figure 3 together with the output of block 305, namely the denoised log Mel spectrum of the current noisy sound signal, for reconstructing the enhanced sound signal (referred to as denoised waveform in figure 3).

[0042] As will be appreciated from figures 2 and 3, in the application phase, the neural network 103 and the processing circuitry 101 are configured to process the current noisy sound signal into the enhanced sound signal using only the log Mel spectrum, i.e. the first set of features, of the current noisy sound signal, but not the MFCC coefficients extracted therefrom.

[0043] Figure 4 is a flow diagram showing an example of a corresponding sound processing method 400 according to an embodiment of the invention. The method 400 comprises the steps of: providing 401 the neural network 103, wherein, in a training phase, the neural network 103 has been trained on the basis of the plurality of training signals, including the plurality of training sound signals and the plurality of corresponding training target signals, using a first set of features based on the plurality of training signals and a second set of features based on the plurality of training signals; and, in an application phase, processing 405 the sound signal into the enhanced sound signal using the only the first set of features, but not the second set of features.

[0044] The person skilled in the art will understand that the "blocks" ("units") of the various figures (method and apparatus) represent or describe functionalities of embodiments of the invention (rather than necessarily individual "units" in hardware or software) and thus describe equally functions or features of apparatus embodiments as well as method embodiments (unit = step).

[0045] In the several embodiments provided in the present application, it should be understood that the disclosed system, apparatus, and method may be implemented in other manners. For example, the described apparatus embodiment is merely exemplary. For example, the unit division is merely logical function division and may be other division in actual implementation. For example, a plurality of units or components may be combined or integrated into another system, or some features may be ignored or not performed. In addition, the displayed or discussed mutual couplings or direct couplings or communication connections may be implemented by using some interfaces. The indirect couplings or communication connections between the apparatuses or units may be implemented in electronic, mechanical, or other forms.

[0046] The units described as separate parts may or may not be physically separate, and parts displayed as units may or may not be physical units, may be located in one position, or may be distributed on a plurality of network units. Some or all of the units may be selected according to actual needs to achieve the objectives of the solutions of the embodiments.

[0047] In addition, functional units in the embodiments of the present invention may be integrated into one processing unit, or each of the units may exist alone physically, or two or more units are integrated into one unit.

Claims

1. A sound processing apparatus (100) configured to process a sound signal comprising a target signal and a current noise signal into an enhanced sound

signal, wherein the apparatus (100) comprises:

processing circuitry (101) configured to provide a neural network (103), wherein the neural network (103), has been trained on the basis of a plurality of training signals, including a plurality of training sound signals and a plurality of training target signals, using a first set of features based on the plurality of training signals and a second set of features based on the plurality of training signals, and
wherein processing circuitry (101) is configured to process the sound signal into the enhanced sound signal using the first set of features, but not the second set of features.

2. The apparatus (100) of claim 1, wherein the processing circuitry (101) is configured to transform the plurality of training signals from the time domain into the frequency domain and wherein the neural network (103), in the training phase, has been trained on the basis of the plurality of training signals in the frequency domain.

3. The apparatus (100) of claim 2, wherein, in the training phase, the neural network (103) is further configured to generate for each of the plurality training sound signals an estimated training target signal and to generate the first set of features and the second set of features on the basis of the plurality of estimated training target signals and the plurality of training target signals.

4. The apparatus (100) of claim 3, wherein, in the training phase, the neural network (103) is further configured to generate the first set of features on the basis of the plurality of estimated training target signals and the plurality of training target signals and the second set of features on the basis of the first set of features.

5. The apparatus (100) of claim 4, wherein the first set of features comprises a plurality of spectra, in particular log mel spectra of the plurality of estimated training target signals and a plurality of spectra, in particular log mel spectra of the plurality of training target signals.

6. The apparatus (100) of claim 5, wherein the second set of features comprises a first plurality of MFCC coefficients generated on the basis of the plurality of spectra of the plurality of estimated training target signals and a second plurality of MFCC coefficients generated on the basis of the plurality of spectra of the plurality of training target signals.

7. The apparatus (100) of claim 6, wherein the neural network (103) has been trained by minimizing a first difference measure between the plurality of spectra, in particular log mel spectra of the plurality of estimated training target signals and the plurality of spectra, in particular log mel spectra of the plurality of training target signals and a second difference measure between the first plurality of MFCC coefficients and the second plurality of MFCC coefficients.

8. The apparatus (100) of any one of claims 3 to 7, wherein the processing circuitry (101) is configured to generate for each of the plurality of training sound signals a two-dimensional data array, wherein each row of the two-dimensional data array comprises a spectrum of the respective training sound signal, wherein the neural network (103) is configured to generate the estimated training target signal for the respective training sound signal by removing noise from the two-dimensional data array.

9. The apparatus (100) of claim 8, wherein the neural network (103) is configured to remove the noise from the two-dimensional data array using image processing.

10. The apparatus (100) of any one of the preceding claims, wherein the neural network (103) comprises one or more convolutional layers and deconvolutional layers and/or one or more skip connections.

11. A method (400) for processing a sound signal comprising a target signal and a current noise signal into an enhanced sound signal, wherein the method (401) comprises:

providing (401) a neural network (103), wherein, in a training phase, the neural network (103) has been trained on the basis of a plurality of training signals, including a plurality of training sound signals and a plurality of training target signals, using a first set of features based on the plurality of training signals and a second set of features based on the plurality of training signals; and
in an application phase, processing (403) the sound signal into the enhanced sound signal using the first set of features but not the second set of features.

12. A computer program product comprising program code configured to control a sound processing apparatus to perform the method (400) according to claim 11 when executed on a computer or processor.

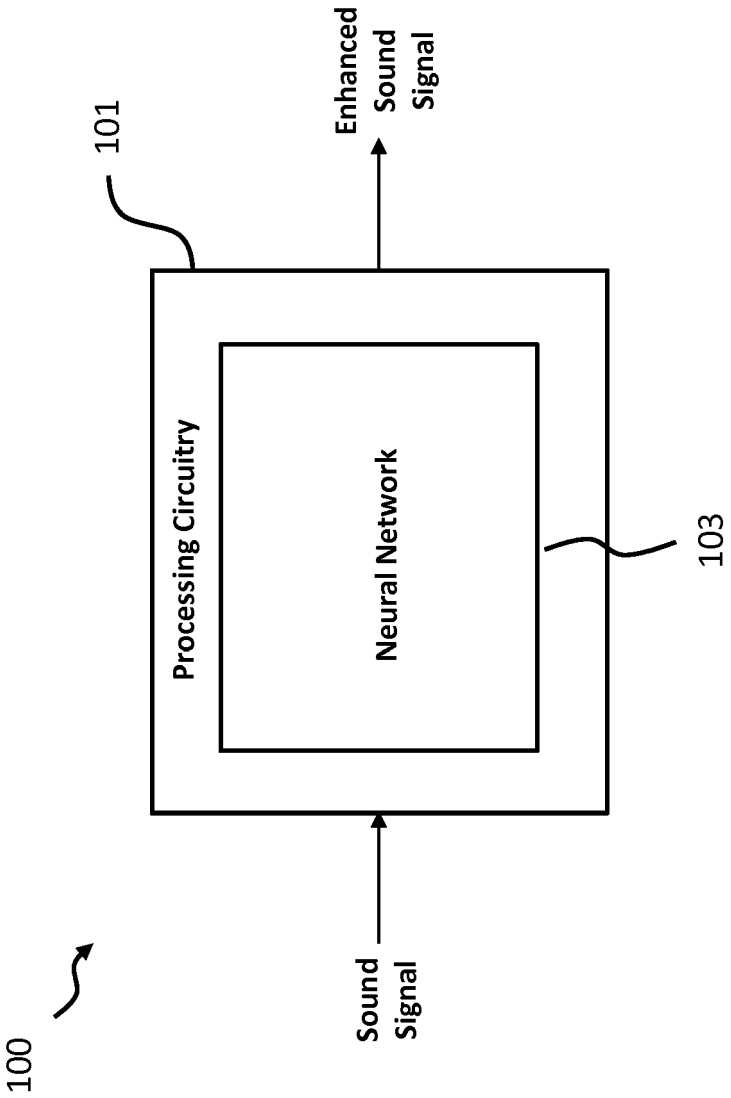

100
101
Sound Signal
Processing Circuitry
Neural Network
103
Enhanced Sound Signal

Fig. 1

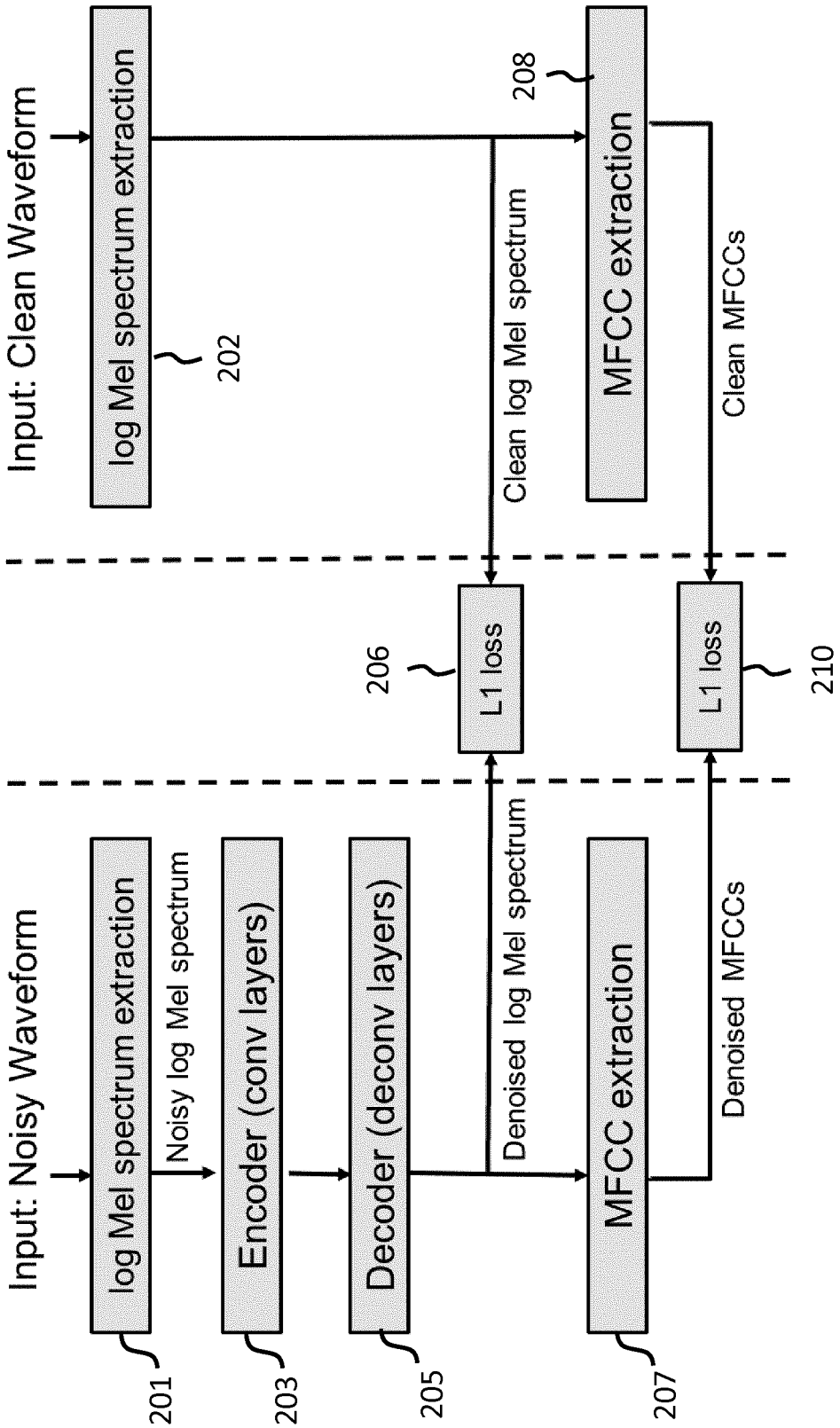

Input: Noisy Waveform
log Mel spectrum extraction
201
Noisy log Mel spectrum
Encoder (conv layers)
203
Decoder (deconv layers)
205
Denoised log Mel spectrum
MFCC extraction
207
Denoised MFCCs
L1 loss
206
Input: Clean Waveform
log Mel spectrum extraction
202
Clean log Mel spectrum
MFCC extraction
208
Clean MFCCs
L1 loss
210

Fig. 2

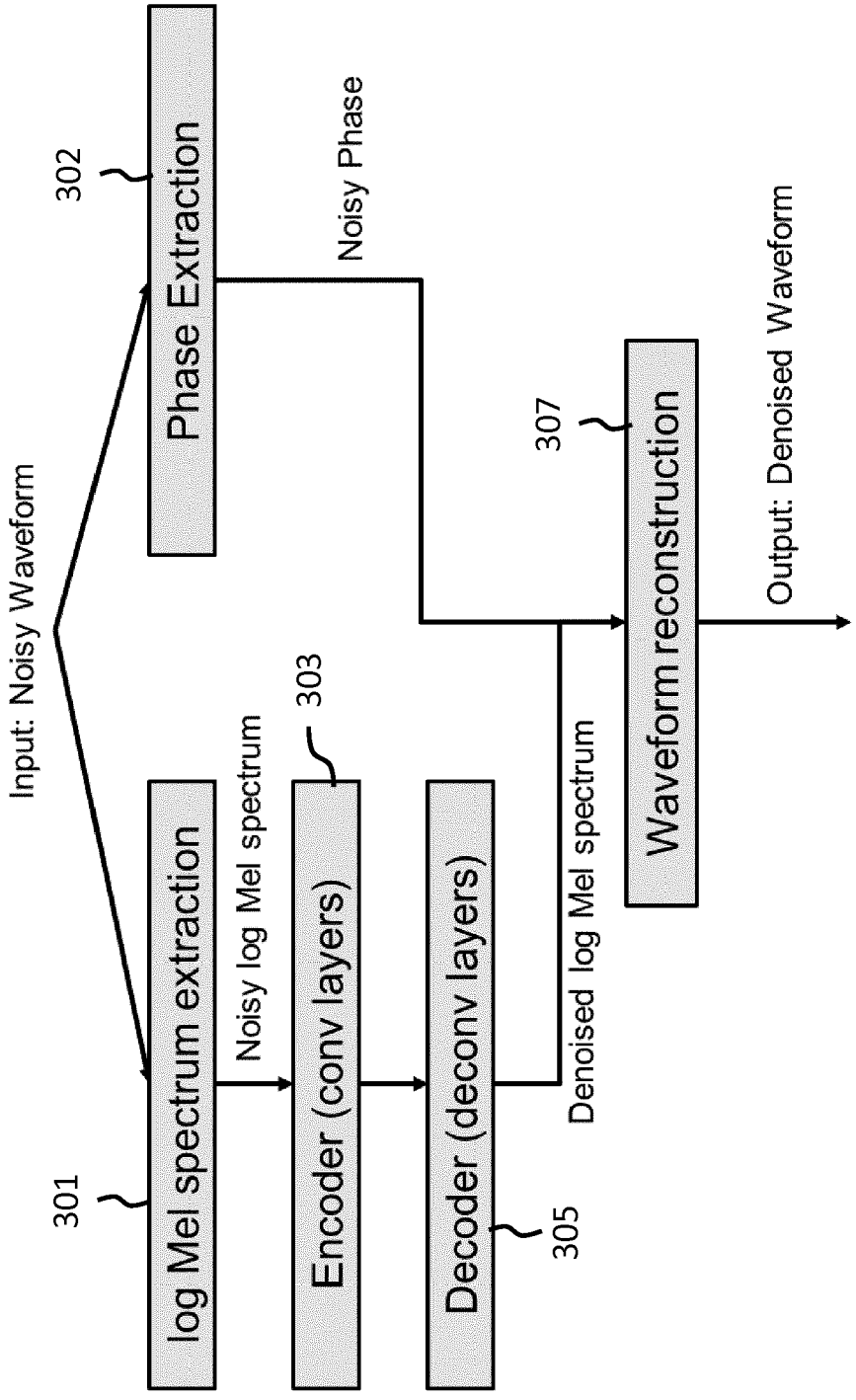

Input: Noisy Waveform
302
Phase Extraction
Noisy Phase
307
Waveform reconstruction
Output: Denoised Waveform
303
Noisy log Mel spectrum
Encoder (conv layers)
Decoder (deconv layers)
Denoised log Mel spectrum
301
log Mel spectrum extraction
305

Fig. 3

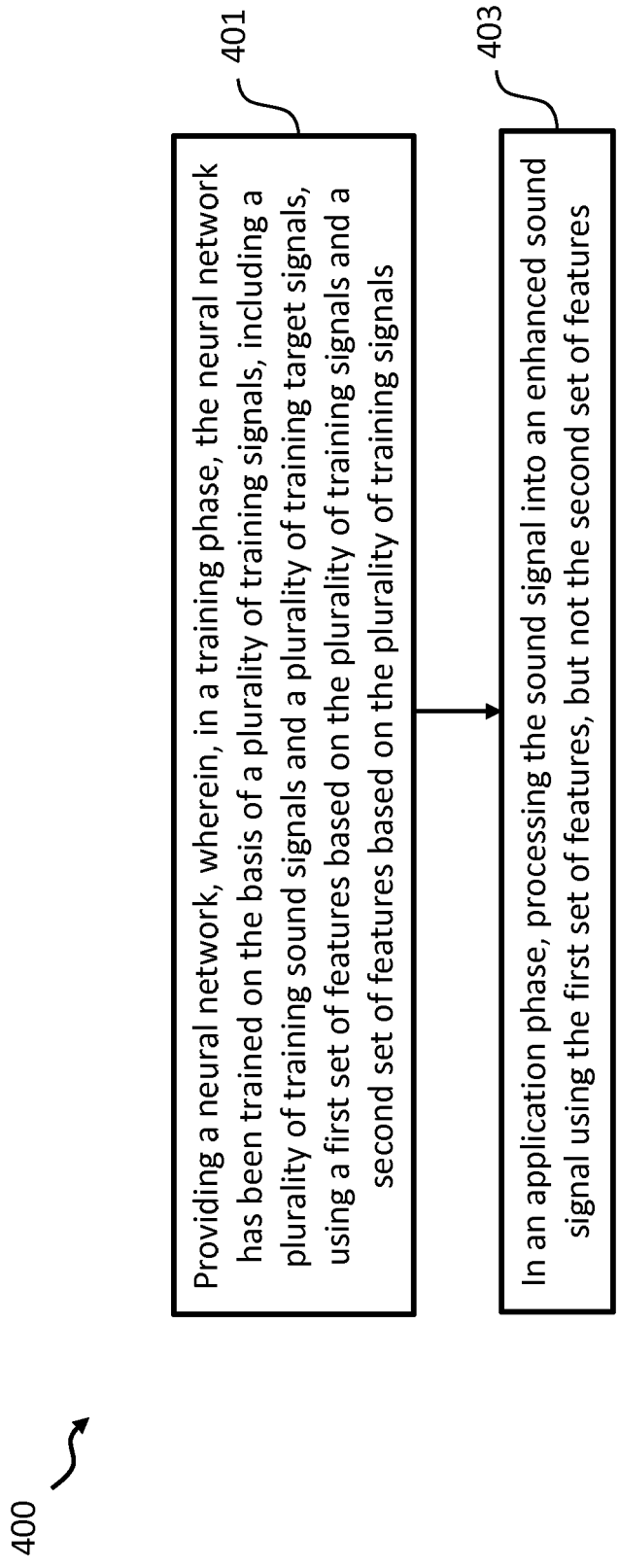

400
401
Providing a neural network, wherein, in a training phase, the neural network has been trained on the basis of a plurality of training signals, including a plurality of training sound signals and a plurality of training target signals, using a first set of features based on the plurality of training signals and a second set of features based on the plurality of training signals
403
In an application phase, processing the sound signal into an enhanced sound signal using the first set of features, but not the second set of features

Fig. 4

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 19 15 0642

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
X	SE RIM PARK ET AL: "A Fully Convolutional Neural Network for Speech Enhancement", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 22 September 2016 (2016-09-22), XP080813105,	1,2, 10-12	INV. G10L21/0208 G10L25/30 G10L21/0232 G10L21/0264
Y	* figures 1,3 * * sections 2, 3.2, 3.3 and 4.1 *	3-9	
Y	WENINGER FELIX ET AL: "Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition", 2014 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING (ICASSP), IEEE, 4 May 2014 (2014-05-04), pages 4623-4627, XP032617061, DOI: 10.1109/ICASSP.2014.6854478 [retrieved on 2014-07-11] * figure 1 * * sections 3.2 and 6 *	3-9	TECHNICAL FIELDS SEARCHED (IPC) G10L
A	ZIXING ZHANG ET AL: "Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 30 May 2017 (2017-05-30), XP081072447, * figure 1 * * section III *	1-12	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
Munich	12 June 2019	Tilp, Jan

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Non-patent literature cited in the description

- **LONG ; SHELHAMER.** *Fully Convolutional Networks for Semantic Segmentation,* 2015 **[0039]**