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

(11) **EP 3 797 415 B1**

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention of the grant of the patent:
19.06.2024 Bulletin 2024/25

(21) Application number: **18752715.5**

(22) Date of filing: **02.08.2018**

(51) International Patent Classification (IPC):
G10L 21/0208 *(2013.01)* ***G10L 25/30*** *(2013.01)*
G10L 21/0232 *(2013.01)*

(52) Cooperative Patent Classification (CPC):
G10L 21/0208; G10L 21/0232; G10L 25/30

(86) International application number:
PCT/EP2018/071070

(87) International publication number:
WO 2020/025140 (06.02.2020 Gazette 2020/06)

(54) SOUND PROCESSING APPARATUS AND METHOD FOR SOUND ENHANCEMENT

KLANGVERARBEITUNGSVORRICHTUNG UND VERFAHREN ZUR KLANGVERBESSERUNG

APPAREIL DE TRAITEMENT SONORE ET PROCÉDÉ D'AMÉLIORATION SONORE

(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

(43) Date of publication of application:
31.03.2021 Bulletin 2021/13

(73) Proprietor: **HUAWEI TECHNOLOGIES CO., LTD.**
Shenzhen, Guangdong 518129 (CN)

(72) Inventors:
- **KEREN, Gil**
  80992 Munich (DE)
- **HAN, Jing**
  Shenzhen, Guangdong 518129 (CN)
- **SCHULLER, Bjoern**
  Shenzhen, Guangdong 518129 (CN)
- **JIN, Wenyu**
  Shenzhen, Guangdong 518129 (CN)
- **SETIAWAN, Panji**
  80992 Munich (DE)
- **GROSCHE, Peter**
  80992 Munich (DE)

(74) Representative: **MERH-IP Matias Erny Reichl Hoffmann**
Patentanwälte PartG mbB
Paul-Heyse-Strasse 29
80336 München (DE)

(56) References cited:
US-A1- 2018 040 333

- **ANURAG KUMAR ET AL: "Speech Enhancement in Multiple-Noise Conditions Using Deep Neural Networks", INTERSPEECH 2016, vol. 2016, 12 September 2016 (2016-09-12), pages 3738-3742, XP55567333, ISSN: 1990-9772, DOI: 10.21437/Interspeech.2016-88**
- **YONG XU ET AL: "Dynamic Noise Aware Training for Speech Enhancement Based on Deep Neural Networks", INTERSPEECH 2014, 14 September 2014 (2014-09-14), pages 2670-2674, XP055576602, Singapore**
- **CHOI J ET AL: "An auditory-based adaptive speech enhancement system by neural network according to noise intensity", CIRCUITS AND SYSTEMS, 2000. 42ND MIDWEST SYMPOSIUM ON AUGUST 8 - 11, 1999, PISCATAWAY, NJ, USA,IEEE, vol. 2, 8 August 1999 (1999-08-08), pages 993-996, XP010511117, ISBN: 978-0-7803-5491-3**

EP 3 797 415 B1

Description

TECHNICAL FIELD

[0001] The invention relates to the field of sound processing. More specifically, the invention relates to a sound processing apparatus and method for sound, in particular speech enhancement.

BACKGROUND

[0002] Sound or audio enhancement conventionally uses only a recording of the speech and environment, i.e. noise for producing the enhanced speech audio. Often audio enhancement procedures make use of neural network, such as the speech enhancement procedure described in the article "A Fully Convolutional Neural Network For Speech Enhancement", Se Rim Park and Jinwon Lee, in Proc. Interspeech 2017, August 20-24, 2017, pages 1993-1997, Stockholm, Sweden.

[0003] However, given only one recording that contains both the speech and the noise created by the environment, it can be difficult, in particular for a neural network to ascertain which components of an audio signal originate from the environment, which components are the clean speech or sound, i.e. the target signal and which components are just reverberation effects of both the speech and the environment. Additionally, in multichannel settings, audio localization can be performed, but sound enhancement may have difficulties predicting whether a given sound source is to be attributed to the speech or the environment.

[0004] Thus, there is still a need for an improved sound processing apparatus and method allowing for an improved enhancement of a noisy sound signal.

[0005] In Anurag, Kumar et al., 'Speech Enhancement in Multiple Noise Conditions Using Deep Neural Networks' (Interspeech 2016, Vol. 2016, (2016), p. 3738 - 3742) Deep Neuronal Networks training strategies based on psychoacoustic models from speech coding were used for speech enhancement.

[0006] Similarly, in Yong Xu et al., 'Dynamic Noise Aware Training for Speech Enhancement Based on Deep Neural Networks' (Interspeech 2014, (2014), p. 2670 - 2674) Deep Neural Network algorithm based on noise aware training by incorporating noise information, noise type generalization enrichments and global variance equalization is proposed.

[0007] Furthermore, in Choi, J. et al, 'An auditory-based adaptive speech enhancement system by neural network according to noise intensity' (Circuits and Systems, Vol. 2, (200), p. 993 - 996) an additional speech enhancement strategy called lateral inhibition is proposed, capable of estimating noise intensities with the aid of neural networks.

[0008] Additionally, in US 2018/040333 A1 a method for performing speech enhancement using Deep Neural Networks is suggested that starts training a microphone by target training signals including signal approximation of clean speech. Moreover, estimations of loudspeaker signals are performed based on Acoustic-echo-cancelling signals and the aforementioned microphone signal.

SUMMARY

[0009] It is an object of the invention to provide an improved sound processing apparatus and method allowing for an improved enhancement of a noisy sound signal.

[0010] 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.

[0011] Generally, embodiments of the invention are based on the idea to use for a plurality of training sound signals, including a training target signal and a training noise signal, the training noise signal as an additional input for training the neural network of a sound processing apparatus for improving the sound enhancement process. In an embodiment, the environment recording, i.e. the training noise signal can be fed into a dedicated portion of the neural network that outputs an audio environment representation defined, for instance, by a parameter set. The environment representation, in turn, can be fed as an additional input to another portion of the neural network that produces the enhanced sound. By explicitly learning to represent audio environments, and use these representations for enhancement, embodiments of the invention allow to perform efficient speech enhancement in unpredictable audio environments.

[0012] The invention can be implemented in hardware and/or software.

BRIEF DESCRIPTION OF THE DRAWINGS

[0013] Further embodiments of the invention will be described with respect to the following figures, wherein:

Fig. 1a shows a schematic diagram illustrating an example of processing blocks implemented in a single channel sound processing apparatus according to an embodiment in a training phase;

Fig. 1b shows a schematic diagram illustrating an example of processing blocks implemented in a single channel sound processing apparatus according to an embodiment in an application phase;

Fig. 2a shows a schematic diagram illustrating an example of processing blocks implemented in a multi-channel sound processing apparatus according to an embodiment in a training phase;

Fig. 2b shows a schematic diagram illustrating an example of processing blocks implemented in a multi-channel sound processing apparatus according to an embodiment in an application phase; and

Fig. 3 shows a flow diagram illustrating an example of a sound processing method according to an embodiment.

[0014] In the various figures, identical reference signs will be used for identical or at least functionally equivalent features.

DETAILED DESCRIPTION OF EMBODIMENTS

[0015] In the following description, reference is made to the accompanying drawings, which form part of the disclosure, and in which are shown, by way of illustration, specific aspects in which the invention may be placed. It is understood that other aspects may be utilized and structural or logical changes may be made without departing from the scope of the invention. The following detailed description, therefore, is not to be taken in a limiting sense, as the scope of the invention is defined by the appended claims.

[0016] For instance, it is 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 a specific method step is described, a corresponding device may include a unit to perform the described method step, even if such unit is not explicitly described or illustrated in the figures. Further, it is understood that the features of the various exemplary aspects described herein may be combined with each other, unless specifically noted otherwise.

[0017] Figure 1a shows a schematic diagram illustrating an example of processing blocks implemented in a single channel sound processing apparatus 100 according to an embodiment in a training phase, while figure 1b shows a schematic diagram illustrating an example of processing blocks implemented in the single channel sound processing apparatus 100 in an application phase.

[0018] As will be described in more detail further below, the sound processing apparatus 100 is configured to process a current noisy sound, in particular speech signal comprising a target signal and a current noise signal into an enhanced, i.e. de-noised sound, in particular speech signal.

[0019] The apparatus 100, which could be implemented, for instance, as a loudspeaker, a mobile phone and the like, comprises processing circuitry, in particular one or more processors, configured to provide, i.e. implement an adjustable neural network. In the embodiment shown in figures 1a and 1b, the adjustable neural network comprises a first neural sub-network 103 and a second neural sub-network 107. In an embodiment, the first neural sub-network 103 and/or the second neural sub-network 107 (referred to as "Environment Residual Blocks" 103, 107 in the figures) can comprise one or more residual blocks. In further embodiments, the first neural sub-network 103 and the second neural sub-network 107 can constitute independent, i.e. separate neural networks. In an embodiment, the neural network, the first neural sub-network 103 and/or the second neural sub-network 107 can comprise one or more convolutional layers. More details about possible implementations of the neural network, the first neural sub-network 103 and/or the second neural sub-network 107 can be found, for instance, in the article "A Fully Convolutional Neural Network For Speech Enhancement", Se Rim Park and Jinwon Lee, in Proc. Interspeech 2017, August 20-24, 2017, pages 1993-1997, Stockholm, Sweden.

[0020] In a training phase, the adjustable neural network 103, 107 of the sound processing apparatus 100 is configured to be trained, i.e. conditioned using as a first input a training noise signal (referred to in figure 1a as "Environment Waveform"), as a second input a noisy training sound signal (referred to in figure 1a as "Environment + speech Waveform") comprising a training target signal and the training noise signal and as a third input the training target signal (referred to in figure 1a as "clean Waveform"). Usually, the training phase involves processing a set of training sound signals comprising a plurality of known training target signals and a plurality of known training noise signals.

[0021] In an application phase, the adjustable neural network 103, 107 of the sound processing apparatus 100 is configured to adjust itself on the basis of the current noise signal and to generate an estimated noise signal on the basis of the sound signal comprising the target signal and the current noise signal. The processing circuitry of the sound processing apparatus is further configured to process the sound signal into the enhanced sound signal on the basis of the estimated noise signal.

[0022] As illustrated by blocks 101, 105 and 113 in figures 1a and 1b, in an embodiment, the processing unit of the sound processing apparatus 100 is configured to transform the training noise signal, the noisy training sound signal, the training target signal, the current noise signal and the current sound signal from the time domain into the frequency domain by generating a respective log spectrum thereof. To this end, the blocks 101, 105 and 113 can be configured to perform a short time Fourier transform (STFT) using, for instance, 25 ms frames shifted by 10 ms to extract the spectrum of each signal.

[0023] The spectrum of the training noise signal (which is provided by block 101 of figure 1) is then processed by the first neural sub-network 103. In an embodiment, the first neural sub-network 103 comprises a sequence of residual blocks. In an embodiment, a respective residual block comprises two parallel paths. The first path can contain two convolutional layers applied one after another, where batch normalization and a rectifiedlinear non-linearity are applied in between the layers. The second path can contain the identity function. The respective outputs of the two paths can be summed, and a rectifiedlinear non-linearity can be applied.

[0024] The output provided by the first neural sub-network 103 is a representation of the environment associated with a respective training noise signal (referred to

as "Environment Embedding" in the figures). Thus, in an embodiment, in the training phase (illustrated in figure 1a), the first neural sub-network 103 is configured to generate on the basis of the training noise signal provided by block 101 a parameter set, i.e. an environment embedding vector describing the training noise signal and to provide the parameter set to the second neural sub-network 107, wherein the second neural sub-network 107 is configured to adjust itself on the basis of the parameter set provided by the first neural sub-network 103. Likewise, in the application phase (illustrated in figure 1b), the first neural sub-network 103 is configured to generate on the basis of the current noise signal the environment embedding vector describing the current noise signal and to provide the environment embedding vector to the second neural sub-network 107, wherein the second neural sub-network 107 is configured to adjust itself on the basis of the parameter set provided by the first neural sub-network 103.

[0025] The output of the first neural sub-network 103, i.e. the environment embedding vector describing in the training phase the training noise signal or in the application phase the current noise signal, is used by the second neural sub-network 107 to adjust itself. In other words, the parameter set defined by the environment embedding vector is used as an additional input by the second neural sub-network 107 such that the output of the second neural sub-network 107 depends on the environment embedding vector, and is "adjusting" to the noise in that sense. There can be multiple ways for the second neural sub-network 107 to use this additional input, which also depend on the inner structure of the second neural sub-network 107. In one embodiment, the second neural sub-network 107 comprises a set of residual blocks, each comprised of two convolutional layers. For each convolutional layer, the environment embedding vector is projected (a linear transformation) to a vector with a dimension equal to the number of feature maps in the convolutional layer. Then, the output of this projection is added to every spatial location in the output map of the convolutional layer.

[0026] In the training phase, the adjusted second neural sub-network 107 is configured to generate an estimated training noise signal (referred to as "Enhancement Mask" in figure 1a) on the basis of the training sound signal provided by block 105. Likewise, in the application phase, the adjusted second neural sub-network 107 is configured to generate an estimated noise signal (referred to as "Enhancement Mask" in figure 1a) on the basis of the sound signal provided by block 105.

[0027] In the training phase, in block 109 of the sound processing apparatus 100 shown in figure 1a an enhanced training sound signal (referred to as "Enhanced Speech Spectrum" in figure 1a) is generated on the basis of the estimated training noise signal provided by the second neural sub-network 107 and the training sound signal provided by block 105. In an embodiment, this can be done by subtracting the estimated training noise signal from the training sound signal or, alternatively, by adding the negative of the estimated training noise signal to the training sound signal.

[0028] Likewise, in the application phase, in block 109 of the sound processing apparatus 100 shown in figure 1b an enhanced sound signal (referred to as "Enhanced Speech Spectrum" in figure 1b) is generated on the basis of the estimated noise signal provided by the second neural sub-network 107 and the sound signal provided by block 105. In an embodiment, this can be done by subtracting the estimated noise signal from the sound signal or, alternatively, by adding the negative of the estimated noise signal to the sound signal.

[0029] In the training phase shown in figure 1a, the output of block 109, i.e. the enhanced training sound signal, is used for training the second neural sub-network 107 by minimizing a difference measure, such as the absolute difference(s), the squared difference(s) and the like, between the training target signal provided by block 113 and the enhanced training sound signal provided by block 109. In an embodiment, a gradient-based optimization algorithm can be used for training, i.e. optimizing the model parameters of the second neural sub-network 107.

[0030] In the application phase shown in figure 1b, in block 112 (referred to as "Waveform Reconstruction" in figure 1b) the spectrum of the enhanced sound signal provided by block 109 is transformed back into the time domain. To this end, as illustrated by block 114 of figure 1b, the processing circuitry of the sound processing apparatus 100 can be further configured to extract phase information from the sound signal comprising the target signal and the current noise signal and to transform the spectrum of the enhanced sound signal back into the time domain on the basis of the extracted phase information. In the application phase, the final output of the sound processing apparatus 100 is the enhanced, i.e. de-noised sound signal in the time domain (referred to as "Enhanced Waveform").

[0031] Figures 2a and 2b show a further embodiment of the sound processing apparatus 100 shown in figures 1a and 1b. In the embodiment shown in figures 2a and 2b the sound processing apparatus 100 is configured to process multi-channel sound signals. In the following only the main differences between the embodiment of the sound processing apparatus 100 shown in figures 2a and 2b and the embodiment of the sound processing apparatus 100 shown in figures 1a and 1b will be described.

[0032] As can be taken from figure 2b illustrating the application phase, the processing circuitry of the sound processing apparatus 100 can be configured to select a channel of the multichannel sound signal and to process the multi-channel sound signal into the enhanced sound signal on the basis of the estimated noise signal by subtracting (or adding) the estimated noise signal from the selected channel of the multi-channel sound signal. The selected channel could be, for instance, the channel closest to the speaker. The enhanced spectrum is considered

the output of the beamforming procedure in the multichannel setting.

[0033] Moreover, the processing circuitry in block 114 of figure 2b can be configured to select a channel of the multi-channel sound signal and to extract the phase information from the selected channel of the multi-channel sound signal.

[0034] The multiple channels of the noise signal can be used for localizing the sound sources by processing these channels with a time-frequency transformation that is more localized in time, such as a STFT over frames of 10 ms, shifted by 5 ms, or a wavelet transform.

[0035] Figure 3 shows a flow diagram illustrating an example of a corresponding sound processing method 300 according to an embodiment. The method 300 comprises the steps of: providing 301 the adjustable neural network 103, 107; in a training phase 303, training, i.e. conditioning the adjustable neural network 103, 107 using as a first input a training noise signal, as a second input a noisy training sound signal comprising a training target signal and the training noise signal and as a third input the training target signal; and, in an application phase 305, adjusting the neural network 107 on the basis of the current noise signal, generating an estimated noise signal on the basis of the sound signal comprising the target signal and the current noise signal, and processing the sound signal into the enhanced sound signal on the basis of the estimated noise signal.

[0036] While a particular feature or aspect of the disclosure may have been disclosed with respect to only one of several implementations or embodiments, such feature or aspect may be combined with one or more other features or aspects of the other implementations or embodiments as may be desired and advantageous for any given or particular application. Furthermore, to the extent that the terms "include", "have", "with", or other variants thereof are used in either the detailed description or the claims, such terms are intended to be inclusive in a manner similar to the term "comprise". Also, the terms "exemplary", "for example" and "e.g." are merely meant as an example, rather than the best or optimal. The terms "coupled" and "connected", along with derivatives may have been used. It should be understood that these terms may have been used to indicate that two elements cooperate or interact with each other regardless whether they are in direct physical or electrical contact, or they are not in direct contact with each other.

[0037] Although specific aspects have been illustrated and described herein, it will be appreciated by those of ordinary skill in the art that a variety of alternate and/or equivalent implementations may be substituted for the specific aspects shown and described without departing from the scope of the claims.

[0038] Although the elements in the following claims 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.

[0039] Many alternatives, modifications, and variations will be apparent to those skilled in the art in light of the above teachings. Of course, those skilled in the art readily recognize that there are numerous applications of the invention beyond those described herein. While the invention has been described with reference to one or more particular embodiments, those skilled in the art recognize that many changes may be made thereto without departing from the scope of the invention. It is therefore to be understood that within the scope of the appended claims, the invention may be practiced otherwise than as specifically described herein.

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) **characterised in that** it comprises:

   processing circuitry configured to provide an adjustable neural network (103, 107), wherein the adjustable neural network (103, 107) comprises a first neural sub-network (103) and a second neural sub-network (107) and is configured:
   in a training phase, to be trained using as a first input a training noise signal, as a second input a training sound signal comprising a training target signal and the training noise signal and as a third input the training target signal, and
   in an application phase, to adjust itself on the basis of the current noise signal and to generate an estimated noise signal on the basis of the sound signal, wherein, in the application phase, the first neural sub-network (103) is configured to generate on the basis of the current noise signal a parameter set describing the current noise signal and to provide the parameter set to the second neural sub-network (107), wherein the second neural sub-network (107) is configured to adjust on the basis of the parameter set provided by the first neural sub-network (103);
   wherein, in the application phase, the processing circuitry is further configured to process the sound signal into the enhanced sound signal on the basis of the estimated noise signal.

2. The apparatus (100) of claim 1, wherein the processing circuitry is configured to transform the training noise signal, the training sound signal and the training target signal from a time domain into a frequency domain and wherein the adjustable neural network is configured, in the training phase, to be trained using the training noise signal, the training sound signal

and the training target signal in the frequency domain.

3. The apparatus (100) of claim 1 or 2, wherein the processing circuitry is configured to transform the current noise signal and the sound signal from the time domain into the frequency domain, wherein, in the training phase, the adjustable neural network is configured to adjust itself on the basis of the current noise signal in the frequency domain and to generate the estimated noise signal on the basis of the sound signal comprising the target signal and the current noise signal in the frequency domain and wherein the processing circuitry is configured to process the sound signal into the enhanced sound signal in the frequency domain on the basis of the estimated noise signal in the frequency domain.

4. The apparatus (100) of claim 3, wherein the processing circuitry is further configured to transform the enhanced sound signal from the frequency domain into the time domain.

5. The apparatus (100) of any one of the preceding claims, wherein, in the application phase, the processing circuitry is further configured to extract phase information from the sound signal comprising the target signal and the current noise signal and to process the sound signal into the enhanced sound signal on the basis of the estimated noise signal and the extracted phase information.

6. The apparatus (100) of claim 5, wherein the sound signal is a multi-channel sound signal and wherein, in the application phase, the processing circuitry is configured to select a channel of the multichannel sound signal and to extract the phase information from the selected channel of the multi-channel sound signal.

7. The apparatus (100) of any one of the preceding claims, wherein, in the training phase, the neural network is further configured to generate an estimated training noise signal on the basis of the training sound signal comprising the training target signal and the training noise signal, to process the training sound signal into an enhanced training sound signal on the basis of the estimated training noise signal and to be trained by minimizing a difference measure between the training target signal and the enhanced training sound signal.

8. The apparatus (100) of any one of the preceding claims, wherein, in the application phase, the processing circuitry is configured to process the sound signal into the enhanced sound signal on the basis of the estimated noise signal by subtracting the estimated noise signal from the sound signal.

9. The apparatus (100) of any one of the preceding claims, wherein the sound signal is a multichannel sound signal and wherein, in the application phase, the processing circuitry is configured to select a channel of the multi-channel sound signal and to process the multi-channel sound signal into the enhanced sound signal on the basis of the estimated noise signal by subtracting the estimated noise signal from the selected channel of the multi-channel sound signal.

10. The apparatus (100) of any one of the preceding claims, wherein in the training phase, the first neural sub-network (103) is configured to generate on the basis of the training noise signal a parameter set describing the training noise signal and to provide the parameter set to the second neural sub-network (107), wherein the second neural sub-network (107) is configured to adjust on the basis of the parameter set provided by the first neural sub-network (103).

11. The apparatus (100) of claim 1 or 10, wherein the first neural sub-network (103) and/or the second neural sub-network (107) comprises one or more convolutional layers.

12. A sound processing method (300) for processing a sound signal comprising a target signal and a current noise signal into an enhanced sound signal, wherein the method (300) is **characterised by** comprising:

providing (301) an adjustable neural network (103, 107) comprising a first neural sub-network (103) and a second neural sub-network (107);
in a training phase (303), training the adjustable neural network (103, 107) using as a first input a training noise signal, as a second input a training sound signal comprising a training target signal and the training noise signal and as a third input the training target signal; and
in an application phase (305), adjusting the neural network (103, 107) on the basis of the current noise signal, generating an estimated noise signal on the basis of the sound signal comprising the target signal and the current noise signal, and processing the sound signal into the enhanced sound signal on the basis of the estimated noise signal; wherein, in the application phase, the first neural sub-network (103) is generating on the basis of the current noise signal a parameter set describing the current noise signal and providing the parameter set to the second neural sub-network (107) and wherein the second neural sub-network (107) is adjusting on the basis of the parameter set provided by the first neural sub-network (103).

13. A computer program comprising program code for

performing the method (300) of claim 12, when executed on a computer or a processor.

Patentansprüche

1. Klangverarbeitungsvorrichtung (100), die dazu ausgelegt ist, ein Klangsignal, das ein Sollsignal und ein aktuelles Rauschsignal umfasst, in ein verbessertes Klangsignal zu verarbeiten, wobei die Vorrichtung (100) **dadurch gekennzeichnet ist, dass** sie Folgendes umfasst:
   eine Verarbeitungsschaltungsanordnung, die dazu ausgelegt ist, ein anpassbares neuronales Netz (103, 107) bereitzustellen, wobei das anpassbare neuronale Netz (103, 107) ein erstes neuronales Teilnetz (103) und ein zweites neuronales Teilnetz (107) umfasst und konfiguriert ist:

   in einer Trainingsphase unter Verwendung eines Trainingsrauschsignals als erste Eingabe, eines Trainingsklangsignals, das ein Trainingssollsignal und das Trainingsrauschsignal umfasst, als zweite Eingabe und des Trainingssollsignal als dritte Eingabe trainiert zu werden, und in einer Anwendungsphase sich auf Grundlage des aktuellen Rauschsignals anzupassen und ein geschätztes Rauschsignal auf Grundlage des Klangsignals zu erzeugen, wobei in der Anwendungsphase das erste neuronale Teilnetz (103) dazu ausgelegt ist, auf Grundlage des aktuellen Rauschsignals einen Parametersatz zu erzeugen, der das aktuelle Rauschsignal beschreibt, und den Parametersatz dem zweiten neuronalen Teilnetz (107) bereitzustellen, wobei das zweite neuronale Teilnetz (107) dazu ausgelegt ist, sich auf Grundlage des durch das erste neuronale Teilnetz (103) bereitgestellten Parametersatzes anzupassen;
   wobei in der Anwendungsphase die Verarbeitungsschaltungsanordnung ferner dazu ausgelegt ist, das Klangsignal auf Grundlage des geschätzten Rauschsignals in das verbesserte Klangsignal zu verarbeiten.

2. Vorrichtung (100) gemäß Anspruch 1, wobei die Verarbeitungsschaltungsanordnung dazu ausgelegt ist, das Trainingsrauschsignal, das Trainingsklangsignal und das Trainingssollsignal aus einem Zeitbereich in einen Frequenzbereich zu transformieren, und wobei das anpassbare neuronale Netz dazu ausgelegt ist, in der Trainingsphase unter Verwendung des Trainingsrauschsignals, des Trainingsklangsignals und des Trainingssollsignals im Frequenzbereich trainiert zu werden.

3. Vorrichtung (100) gemäß Anspruch 1 oder 2, wobei die Verarbeitungsschaltungsanordnung dazu ausgelegt ist, das aktuelle Rauschsignal und das Klangsignal aus dem Zeitbereich in den Frequenzbereich zu transformieren, wobei in der Trainingsphase das anpassbare neuronale Netz dazu ausgelegt ist, sich auf Grundlage des aktuellen Rauschsignals im Frequenzbereich anzupassen und das geschätzte Rauschsignal auf Grundlage des Klangsignals, das das Sollsignal und das aktuelle Rauschsignal im Frequenzbereich umfasst, zu erzeugen, und wobei die Verarbeitungsschaltungsanordnung dazu ausgelegt ist, das Klangsignal auf Grundlage des geschätzten Rauschsignals im Frequenzbereich in das verbesserte Klangsignal im Frequenzbereich zu verarbeiten.

4. Vorrichtung (100) gemäß Anspruch 3, wobei die Verarbeitungsschaltungsanordnung ferner dazu ausgelegt ist, das verbesserte Klangsignal aus dem Frequenzbereich in den Zeitbereich zu transformieren.

5. Vorrichtung (100) gemäß einem der vorhergehenden Ansprüche, wobei in der Anwendungsphase die Verarbeitungsschaltungsanordnung ferner dazu ausgelegt ist, Phaseninformationen aus dem Klangsignal umfassend das Sollsignal und das aktuelle Rauschsignal zu extrahieren und das Klangsignal auf Grundlage des geschätzten Rauschsignals und der extrahierten Phaseninformationen in das verbesserte Klangsignal zu verarbeiten.

6. Vorrichtung (100) gemäß Anspruch 5, wobei das Klangsignal ein Mehrkanalklangsignal ist und wobei in der Anwendungsphase die Verarbeitungsschaltungsanordnung dazu ausgelegt ist, einen Kanal des Mehrkanalklangsignals auszuwählen und die Phaseninformationen aus dem ausgewählten Kanal des Mehrkanalklangsignals zu extrahieren.

7. Vorrichtung (100) gemäß einem der vorhergehenden Ansprüche, wobei in der Trainingsphase das neuronale Netz ferner dazu ausgelegt ist, ein geschätztes Trainingsrauschsignal auf Grundlage des Trainingsklangsignals zu erzeugen, das das Trainingssollsignal und das Trainingsrauschsignal umfasst, das Trainingsklangsignal in ein verbessertes Trainingsklangsignal auf Grundlage des geschätzten Trainingsrauschsignals zu verarbeiten und durch Minimieren eines Differenzmaßes zwischen dem Trainingssollsignal und dem verbesserten Trainingsklangsignal trainiert zu werden.

8. Vorrichtung (100) gemäß einem der vorhergehenden Ansprüche, wobei in der Anwendungsphase die Verarbeitungsschaltungsanordnung dazu ausgelegt ist, das Klangsignal auf Grundlage des geschätzten Rauschsignals durch Subtrahieren des geschätzten Rauschsignals von dem Klangsignal in das verbesserte Klangsignal zu verarbeiten.

9. Vorrichtung (100) gemäß einem der vorhergehenden Ansprüche, wobei das Klangsignal ein Mehrkanalklangsignal ist und wobei in der Anwendungsphase die Verarbeitungsschaltungsanordnung dazu ausgelegt ist, einen Kanal des Mehrkanalklangsignals auszuwählen und das Mehrkanalklangsignal auf Grundlage des geschätzten Rauschsignals durch Subtrahieren des geschätzten Rauschsignals von dem ausgewählten Kanal des Mehrkanalklangsignals in das verbesserte Klangsignal zu verarbeiten.

10. Vorrichtung (100) gemäß einem der vorhergehenden Ansprüche, wobei in der Trainingsphase das erste neuronale Netz (103) dazu ausgelegt ist, auf Grundlage des Trainingsrauschsignals einen Parametersatz zu erzeugen, der das Trainingsrauschsignal beschreibt, und den Parametersatz dem zweiten neuronalen Teilnetz (107) bereitzustellen, wobei das zweite neuronale Teilnetz (107) dazu ausgelegt ist, sich auf Grundlage des durch das erste neuronale Teilnetz (103) bereitgestellten Parametersatzes anzupassen.

11. Vorrichtung (100) gemäß Anspruch 1 oder 10, wobei das erste neuronale Teilnetz (103) und/oder das zweite neuronale Teilnetz (107) eine oder mehrere Faltungslagen umfasst.

12. Klangverarbeitungsverfahren (300) zum Verarbeiten eines Klangsignals, das ein Sollsignal und ein aktuelles Rauschsignal umfasst, in ein verbessertes Klangsignal, wobei das Verfahren (300) **dadurch gekennzeichnet ist, dass** es Folgendes umfasst:

Bereitstellen (301) eines anpassbaren neuronalen Netzes (103, 107), das ein erstes neuronales Teilnetz (103) und ein zweites neuronales Teilnetz (107) umfasst;
in einer Trainingsphase (303) das Trainieren des anpassbaren neuronalen Netzes (103, 107), wobei als erste Eingabe ein Trainingsrauschsignal, als zweite Eingabe ein Trainingsklangsignal, das ein Trainingssollsignal und das Trainingsrauschsignal umfasst, und als dritte Eingabe das Trainingssollsignal verwendet wird; und
in einer Anwendungsphase (305) das Anpassen des neuronalen Netzes (103, 107) auf Grundlage des aktuellen Rauschsignals, das Erzeugen eines geschätzten Rauschsignals auf Grundlage des Klangsignals, das das Sollsignal und das aktuelle Rauschsignal umfasst, und das Verarbeiten des Klangsignals in das verbesserte Klangsignal auf Grundlage des geschätzten Rauschsignals; wobei in der Anwendungsphase das erste neuronale Teilnetz (103) auf Grundlage des aktuellen Rauschsignals einen Parametersatz erzeugt, der das aktuelle Rauschsignal beschreibt, und den Parametersatz dem zweiten neuronalen Teilnetz (107) bereitstellt, und wobei sich das zweite neuronale Netz (107) auf Grundlage des durch das erste neuronale Teilnetz (103) bereitgestellten Parametersatzes anpasst.

13. Computerprogramm, umfassend Programmcode zum Durchführen des Verfahrens (300) gemäß Anspruch 12, wenn es auf einem Computer oder einem Prozessor ausgeführt wird.

Revendications

1. Appareil (100) de traitement du son configuré pour traiter un signal sonore comprenant un signal cible et un signal de bruit courant pour obtenir un signal sonore amélioré, l'appareil (100) étant **caractérisé en ce qu'**il comprend :
une circuiterie de traitement configurée pour fournir un réseau neuronal ajustable (103, 107), le réseau neuronal ajustable (103, 107) comprenant un premier sous-réseau neuronal (103) et un deuxième sous-réseau neuronal (107) et étant configuré :

dans une phase d'entraînement, pour être entraîné en utilisant, comme première entrée, un signal de bruit d'entraînement, comme deuxième entrée, un signal sonore d'entraînement comprenant un signal cible d'entraînement et le signal de bruit d'entraînement et, comme troisième entrée, le signal cible d'entraînement, et dans une phase d'application, pour s'ajuster sur la base du signal de bruit courant et pour générer un signal de bruit estimé sur la base du signal sonore ; dans la phase d'application, le premier sous-réseau neuronal (103) étant configuré pour générer, sur la base du signal de bruit courant, un ensemble de paramètres décrivant le signal de bruit courant et pour fournir l'ensemble de paramètres au deuxième sous-réseau neuronal (107), le deuxième sous-réseau neuronal (107) étant configuré pour s'ajuster sur la base de l'ensemble de paramètres fourni par le premier sous-réseau neuronal (103) ;
dans la phase d'application, la circuiterie de traitement étant configurée en outre pour traiter le signal sonore pour obtenir le signal sonore amélioré sur la base du signal de bruit estimé.

2. Appareil (100) selon la revendication 1, la circuiterie de traitement étant configurée pour transformer le signal de bruit d'entraînement, le signal sonore d'entraînement et le signal cible d'entraînement d'un domaine temporel à un domaine fréquentiel, et le réseau neuronal ajustable étant configuré, dans la

phase d'entraînement, pour être entraîné en utilisant le signal de bruit d'entraînement, le signal sonore d'entraînement et le signal cible d'entraînement dans le domaine fréquentiel.

3. Appareil (100) selon la revendication 1 ou 2, la circuiterie de traitement étant configurée pour transformer le signal de bruit courant et le signal sonore du domaine temporel au domaine fréquentiel ; dans la phase d'entraînement, le réseau neuronal ajustable étant configuré pour s'ajuster sur la base du signal de bruit courant dans le domaine fréquentiel et pour générer le signal de bruit estimé sur la base du signal sonore comprenant le signal cible et le signal de bruit courant dans le domaine fréquentiel, et la circuiterie de traitement étant configurée pour traiter le signal sonore pour obtenir le signal sonore amélioré dans le domaine fréquentiel sur la base du signal de bruit estimé dans le domaine fréquentiel.

4. Appareil (100) selon la revendication 3, la circuiterie de traitement étant configurée en outre pour transformer le signal sonore amélioré du domaine fréquentiel au domaine temporel.

5. Appareil (100) selon l'une quelconque des revendications précédentes, dans la phase d'application, la circuiterie de traitement étant configurée en outre pour extraire des informations de phase du signal sonore comprenant le signal cible et le signal de bruit courant et pour traiter le signal sonore pour obtenir le signal sonore amélioré sur la base du signal de bruit estimé et des informations de phase extraites.

6. Appareil (100) selon la revendication 5, le signal sonore étant un signal sonore multicanal et, dans la phase d'application, la circuiterie de traitement étant configurée pour sélectionner un canal du signal sonore multicanal et pour extraire les informations de phase du canal sélectionné du signal sonore multicanal.

7. Appareil (100) selon l'une quelconque des revendications précédentes, dans la phase d'entraînement, le réseau neuronal étant configuré en outre pour générer un signal de bruit d'entraînement estimé sur la base du signal sonore d'entraînement comprenant le signal cible d'entraînement et le signal de bruit d'entraînement, pour traiter le signal sonore d'entraînement pour obtenir un signal sonore d'entraînement amélioré sur la base du signal de bruit d'entraînement estimé et pour être entraîné par minimisation d'une mesure de différence entre le signal cible d'entraînement et le signal sonore d'entraînement amélioré.

8. Appareil (100) selon l'une quelconque des revendications précédentes, dans la phase d'application, la circuiterie de traitement étant configurée pour traiter le signal sonore pour obtenir le signal sonore amélioré sur la base du signal de bruit estimé par soustraction du signal de bruit estimé au signal sonore.

9. Appareil (100) selon l'une quelconque des revendications précédentes, le signal sonore étant un signal sonore multicanal et, dans la phase d'application, la circuiterie de traitement étant configurée pour sélectionner un canal du signal sonore multicanal et pour traiter le signal sonore multicanal pour obtenir le signal sonore amélioré sur la base du signal de bruit estimé par soustraction du signal de bruit estimé au canal sélectionné du signal sonore multicanal.

10. Appareil (100) selon l'une quelconque des revendications précédentes, dans la phase d'entraînement, le premier sous-réseau neuronal (103) étant configuré pour générer, sur la base du signal de bruit d'entraînement, un ensemble de paramètres décrivant le signal de bruit d'entraînement et pour fournir l'ensemble de paramètres au deuxième sous-réseau neuronal (107), le deuxième sous-réseau neuronal (107) étant configuré pour s'ajuster sur la base de l'ensemble de paramètres fourni par le premier sous-réseau neuronal (103).

11. Appareil (100) selon la revendication 1 ou 10, le premier sous-réseau neuronal (103) et/ou le deuxième sous-réseau neuronal (107) comprenant une ou plusieurs couches convolutives.

12. Procédé (300) de traitement du son pour traiter un signal sonore comprenant un signal cible et un signal de bruit courant pour obtenir un signal sonore amélioré, le procédé (300) étant **caractérisé en ce qu'**il comprend :

la fourniture (301) d'un réseau neuronal ajustable (103, 107) comprenant un premier sous-réseau neuronal (103) et un deuxième sous-réseau neuronal (107) ;
dans une phase d'entraînement (303), l'entraînement du réseau neuronal ajustable (103, 107) en utilisant, comme première entrée, un signal de bruit d'entraînement, comme deuxième entrée, un signal sonore d'entraînement comprenant un signal cible d'entraînement et le signal de bruit d'entraînement et, comme troisième entrée, le signal cible d'entraînement; et
dans une phase d'application (305), l'ajustement du réseau neuronal (103, 107) sur la base du signal de bruit courant, la génération d'un signal de bruit estimé sur la base du signal sonore comprenant le signal cible et le signal de bruit courant, et le traitement du signal sonore pour obtenir le signal sonore amélioré sur la base du signal de bruit estimé ; dans la phase d'ap-

plication, le premier sous-réseau neuronal (103) générant, sur la base du signal de bruit courant, un ensemble de paramètres décrivant le signal de bruit courant et fournissant l'ensemble de paramètres au deuxième sous-réseau neuronal (107), et le deuxième sous-réseau neuronal (107) s'ajustant sur la base de l'ensemble de paramètres fourni par le premier sous-réseau neuronal (103).

13. Programme d'ordinateur comprenant un code de programme pour réaliser le procédé (300) selon la revendication 12, lorsqu'il s'exécute sur un ordinateur ou un processeur.

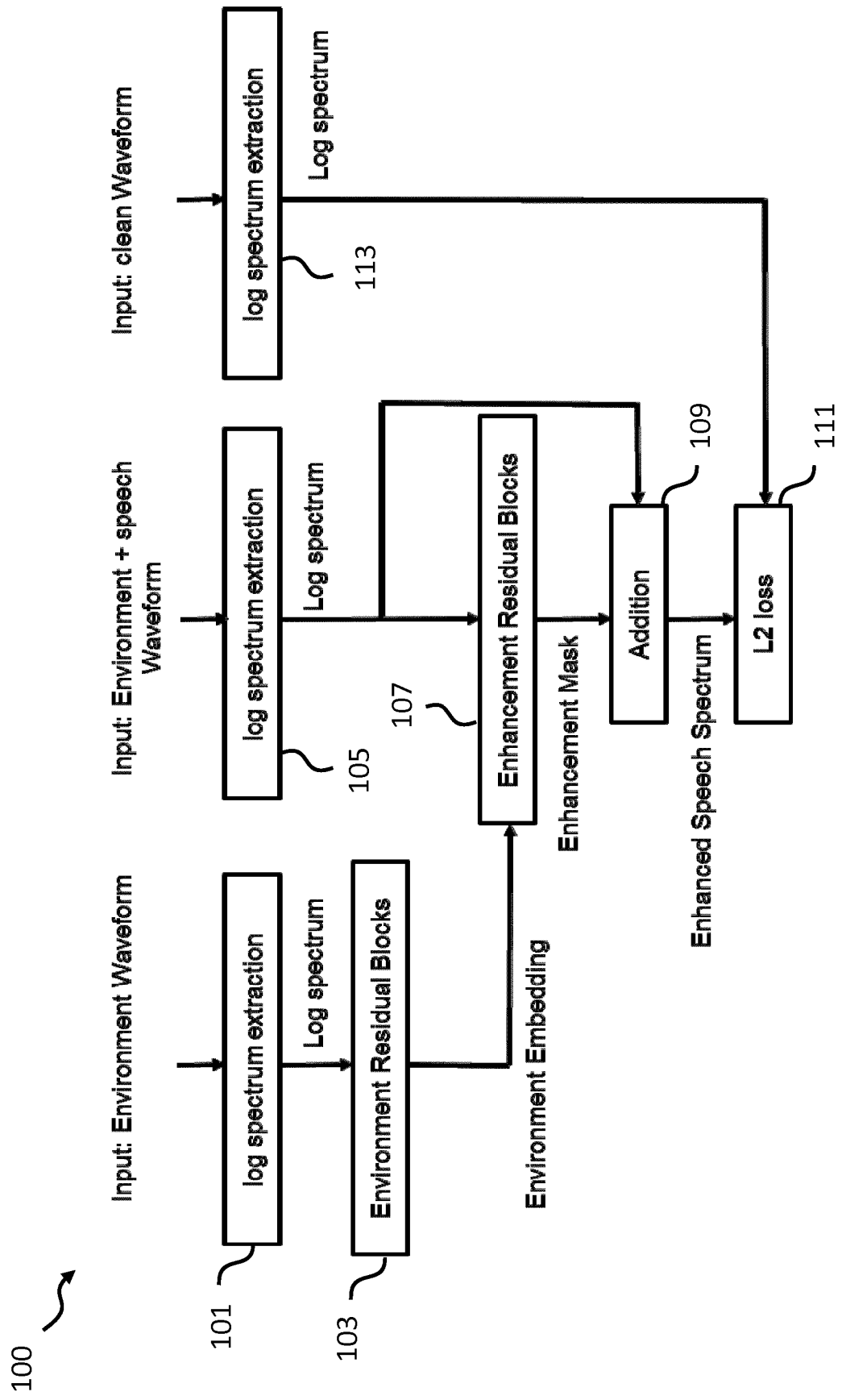
100
Input: Environment Waveform
log spectrum extraction
Log spectrum
Environment Residual Blocks
101
103
Environment Embedding
Input: Environment + speech Waveform
log spectrum extraction
Log spectrum
105
107
Enhancement Residual Blocks
Enhancement Mask
Addition
109
Enhanced Speech Spectrum
L2 loss
111
Input: clean Waveform
log spectrum extraction
Log spectrum
113

Fig. 1a

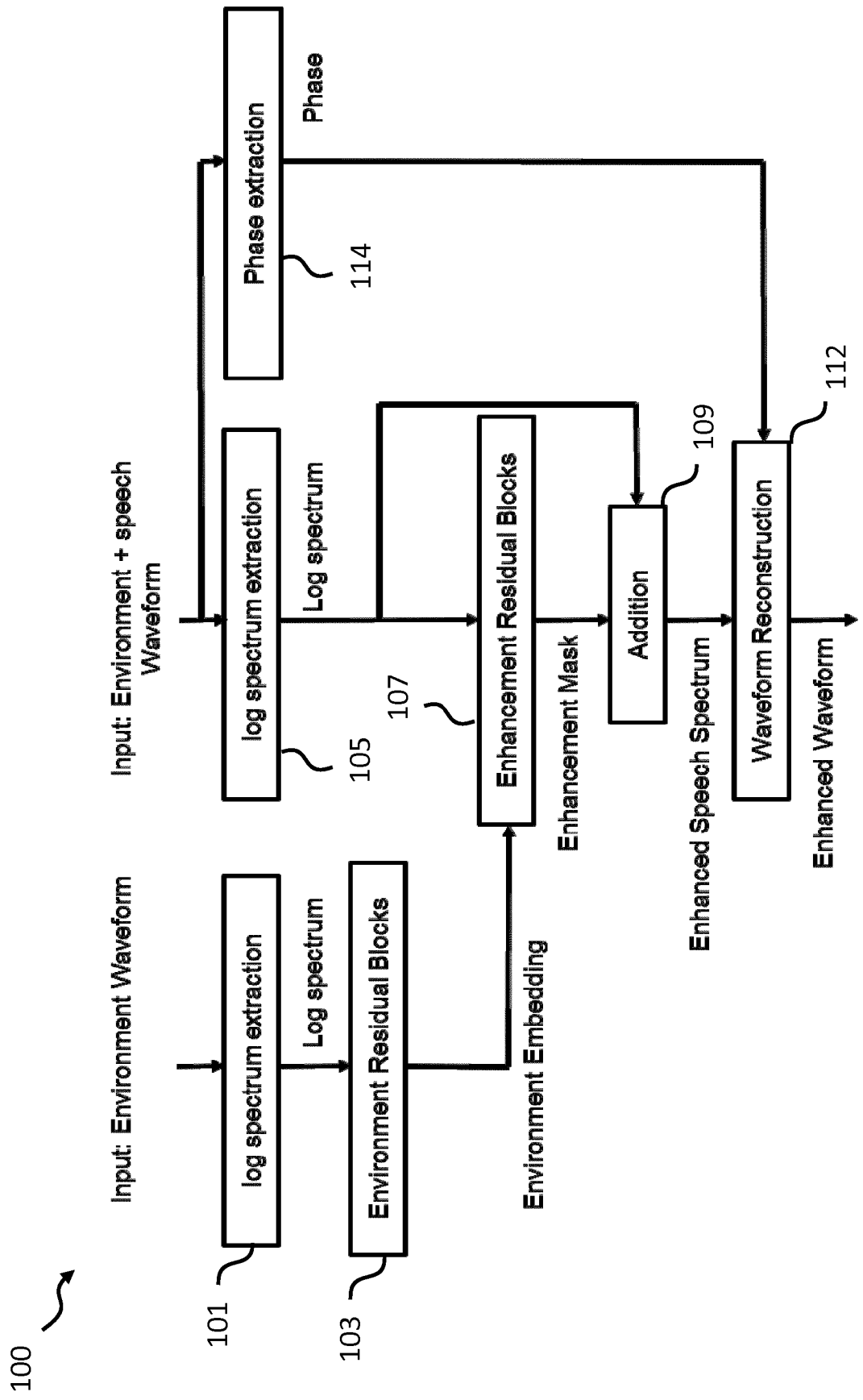
100
Input: Environment Waveform
log spectrum extraction
101
Log spectrum
Environment Residual Blocks
103
Environment Embedding
Input: Environment + speech Waveform
log spectrum extraction
105
Log spectrum
107
Enhancement Residual Blocks
Enhancement Mask
Addition
109
Enhanced Speech Spectrum
Phase extraction
114
Phase
Waveform Reconstruction
112
Enhanced Waveform

Fig. 1b

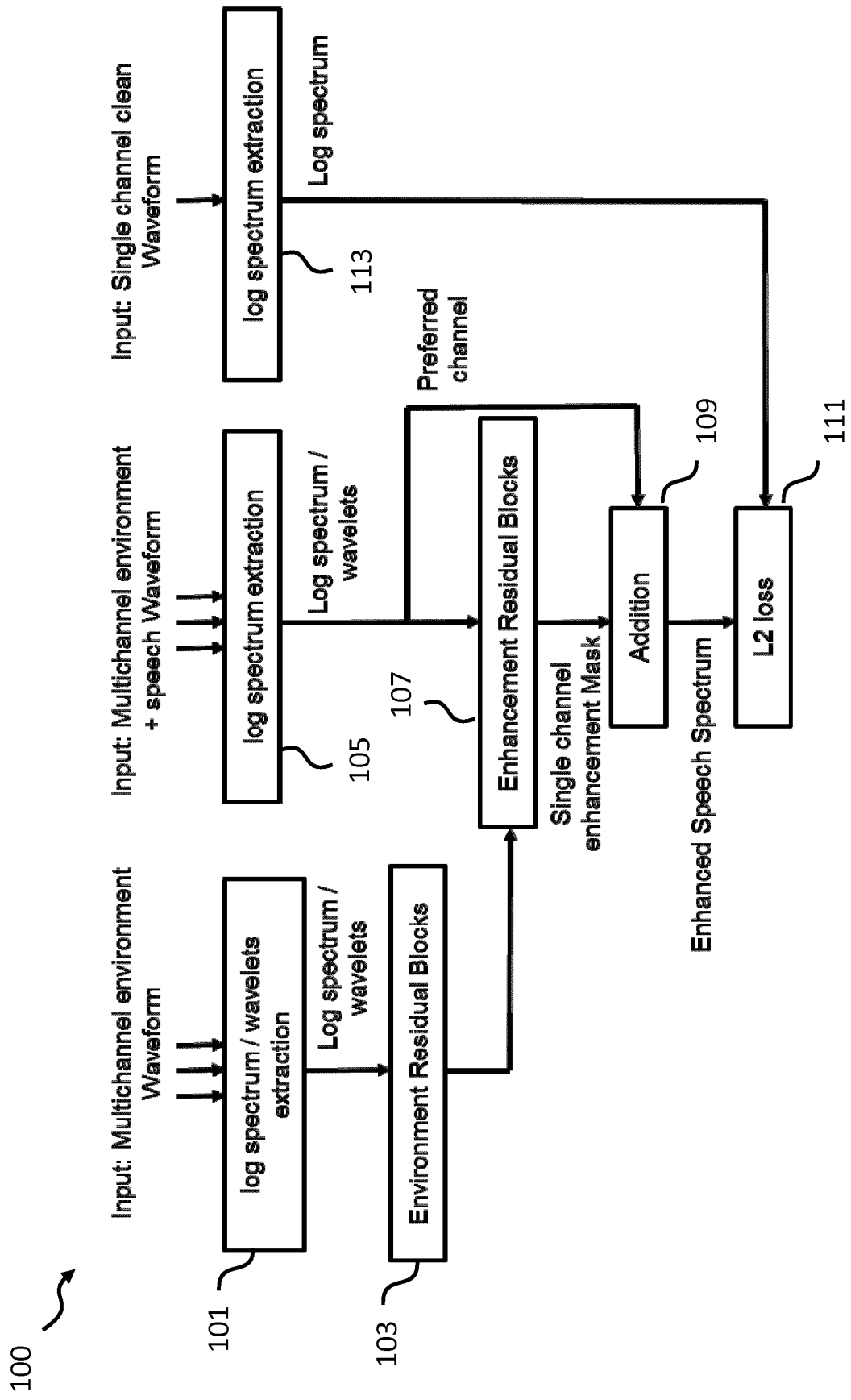
100
Input: Multichannel environment Waveform
log spectrum / wavelets extraction
101
Log spectrum / wavelets
Environment Residual Blocks
103
Input: Multichannel environment + speech Waveform
log spectrum extraction
105
Log spectrum / wavelets
107
Enhancement Residual Blocks
Single channel enhancement Mask
Addition
109
Enhanced Speech Spectrum
L2 loss
111
Input: Single channel clean Waveform
log spectrum extraction
113
Log spectrum
Preferred channel

Fig. 2a

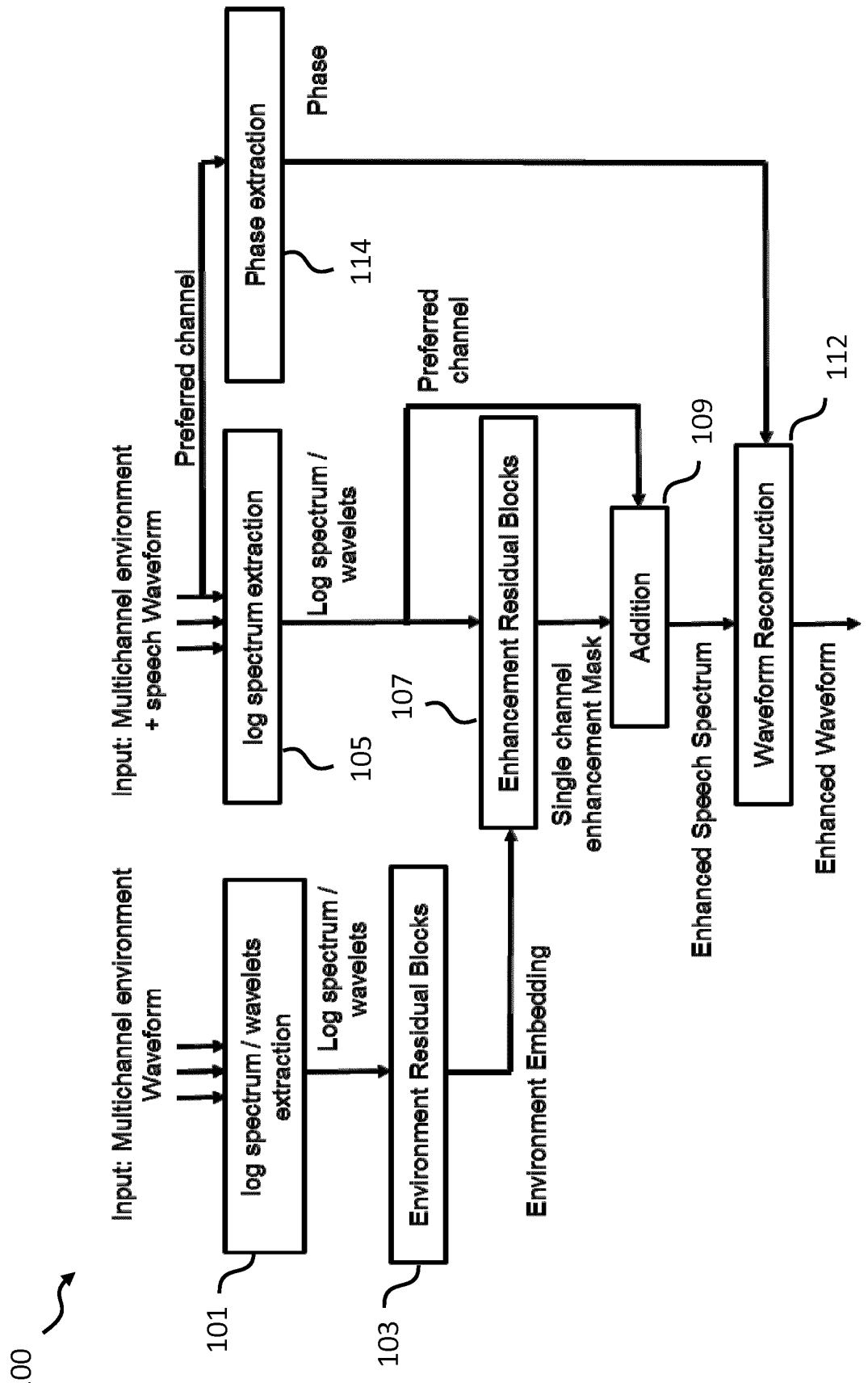
100
Input: Multichannel environment Waveform
log spectrum / wavelets extraction
101
Log spectrum / wavelets
Environment Residual Blocks
103
Environment Embedding
Input: Multichannel environment + speech Waveform
Preferred channel
log spectrum extraction
105
Log spectrum / wavelets
Preferred channel
107
Enhancement Residual Blocks
Single channel enhancement Mask
Addition
109
Enhanced Speech Spectrum
Phase extraction
114
Phase
Waveform Reconstruction
112
Enhanced Waveform

Fig. 2b

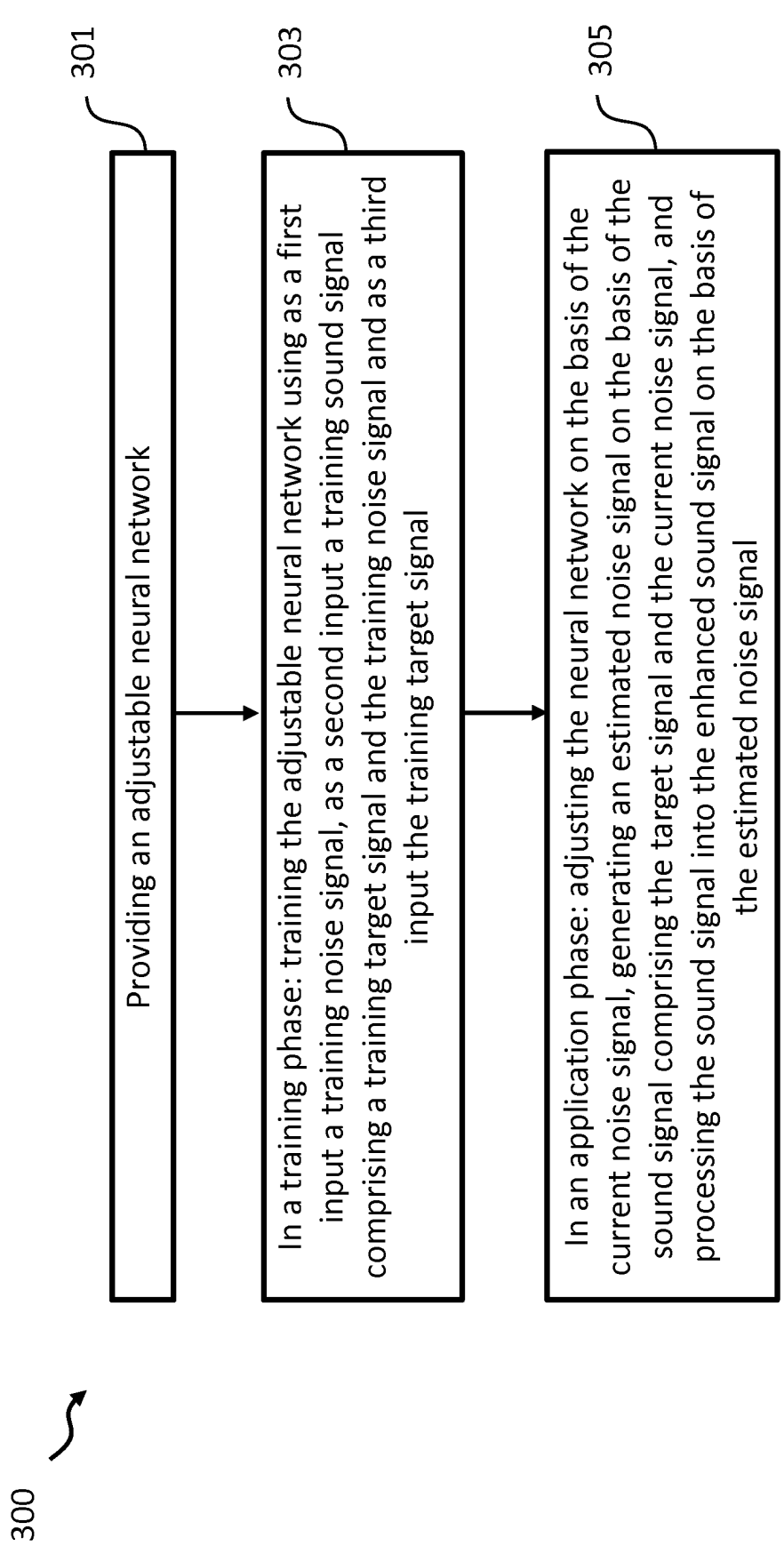
300
301
Providing an adjustable neural network
303
In a training phase: training the adjustable neural network using as a first input a training noise signal, as a second input a training sound signal comprising a training target signal and the training noise signal and as a third input the training target signal
305
In an application phase: adjusting the neural network on the basis of the current noise signal, generating an estimated noise signal on the basis of the sound signal comprising the target signal and the current noise signal, and processing the sound signal into the enhanced sound signal on the basis of the estimated noise signal

Fig. 3

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.

Patent documents cited in the description

- US 2018040333 A1 **[0008]**

Non-patent literature cited in the description

- **SE RIM PARK ; JINWON LEE.** A Fully Convolutional Neural Network For Speech Enhancement. *Proc. Interspeech,* 20 August 2017, vol. 2017, 1993-1997 **[0002] [0019]**
- **ANURAG, KUMAR et al.** Speech Enhancement in Multiple Noise Conditions Using Deep Neural Networks. *Interspeech 2016,* 2016, vol. 2016, 3738-3742 **[0005]**
- **YONG XU et al.** Dynamic Noise Aware Training for Speech Enhancement Based on Deep Neural Networks. *Interspeech,* 2014, vol. 2014, 2670-2674 **[0006]**
- **CHOI, J. et al.** An auditory-based adaptive speech enhancement system by neural network according to noise intensity. *Circuits and Systems,* vol. 2 (200), 993-996 **[0007]**