Claims

1. An apparatus for providing an estimate of a loudness of signal components of interest of an audio signal, wherein the apparatus comprises:

an input interface configured to receive a plurality of samples of the audio signal, and

a neural network configured to receive as input values the plurality of samples of the audio signal or a plurality of derived values being derived from the plurality of samples of the audio signal, and configured to determine at least one output value from the plurality of input values, such that the at least one output value indicates the estimate of the loudness of the signal components of interest of the audio signal.

wherein the audio signal simultaneously comprises the signal components of interest and other signal components of the audio signal, and

wherein an influence of the other signal components on the estimate of the loudness of the signal components of interest is reduced or not present.

2. The apparatus according to claim 1,

wherein the signal components of interest of the audio signal are speech components of the audio signal, and

wherein the neural network is configured to determine the at least one output value from the plurality of input values, such that the at least one output value indicates the estimate of the loudness of the speech components of the audio signal.

3. The apparatus according to claim 2,

Wherein the audio signal simultaneously comprises the speech components and background components of the audio signal, and

wherein an influence of the background components on the estimate of the loudness of the speech components is reduced or not present.

4. The apparatus according to claim 1,

wherein the signal components of interest of the audio signal are sound components of at least one first sound source out of a plurality of sound sources in an environment, and

wherein the audio signal simultaneously comprises the sound components of the at least one first sound source and other sound components of one or more other sound sources out of the plurality of sound sources in the environment, and

wherein the neural network is configured to determine the at least one output value from the plurality of input values, such that the at least one output value indicates the estimate of the loudness of the sound components of the at least one first sound source, and

wherein an influence of the other sound components of the one or more other sound sources on the estimate of the loudness of the sound components of the at least one first sound source is reduced or not present.

5. The apparatus according to claim 4,

wherein the sound components of the at least one first sound source are speech components of a first person out of a plurality of persons speaking in the environment, and

wherein the other sound components of the one or more other sound sources are other speech components of one or more other persons out of the plurality of persons speaking in the environment, and

wherein the audio signal simultaneously comprises the speech components of the first person and the other speech components of the one or more other persons speaking in the environment, and

wherein the neural network is configured to determine the at least one output value from the plurality of input values, such that the at least one output value indicates the estimate of the loudness of the speech components of the first person, and

wherein an influence of the other speech components of the one or more other persons on the estimate of the loudness of the speech components of the first person is reduced or not present.

6. The apparatus according to claim 4,

Wherein the sound components of the at least first sound source are sound components of at least one non-human sound source out of a plurality of non-human sound sources in an environment, and

wherein the other sound components of the one or more other sound sources are other sound components of one or more other non-human sound source out of the plurality of non-human sound sources, and

wherein the audio signal simultaneously comprises the sound components of the at least one first non-human sound source and the other sound components of the one or more other non-human sound sources in the environment, and

wherein the neural network is configured to determine the at least one output value from the plurality of input values, such that the at least one output value indicates the estimate of the loudness of the sound components of the at least one first non-human sound source, and

wherein an influence of the other sound components of the one or more other non-human sound sources on the estimate of the loudness of the sound components of the at least one first non-human sound source is reduced or not present.

7. The apparatus according to claim 4,

wherein the sound components of the at least one first sound source is a singing of one or more singers in the environment, and

wherein the other sound components of the one or more other sound sources are sound components of accompanying musical instruments, which accompany the singing of the one or more singers in the environment, and

wherein the audio signal simultaneously comprises the signing of the one or more singers and the sound components of the accompanying musical instruments, and

wherein the neural network is configured to determine the at least one output value from the plurality of input values, such that the at least one output value indicates the estimate of the loudness of the singing, and

wherein an influence of the sound components of accompanying musical instruments on the estimate of the loudness of the singing is reduced or not present.

8. The apparatus according to one of the preceding claims,

wherein the neural network is configured to determine at least one further output value indicating an estimate of a loudness of the entire audio signal.

9. The apparatus according to one of the preceding claims,

wherein the neural network is configured to determine one or more further output values indicating an estimate of a loudness of the audio signal when speech is present.

10. The apparatus according to one of the preceding claims,

wherein the neural network is configured to determine another one or more output values indicating an estimate of a loudness of background components of the audio signal.

11. The apparatus according to one of the preceding claims,

wherein the apparatus is configured to determine and output at least one other output value indicating an estimate of a partial loudness of the speech components of the audio signal, and

wherein the partial loudness of the speech components of the audio signal depends on the loudness of the speech components of the audio signal and on the loudness of background components of the audio signal.

12. The apparatus according to one of the preceding claims,

wherein the apparatus comprises a postprocessor, configured to modify the estimate of the loudness of the signal components of interest of the audio signal depending on confidence information, and/or configured to output the confidence information, and

wherein the confidence information indicates a reliability on whether or not the estimate of the loudness of the signal components of interest of the audio signal conducted by the neural network is reliable, or wherein the confidence information indicates one or more values indicating a degree of reliability of the estimate of the loudness of the signal components of interest of the audio signal conducted by the neural network.

13. The apparatus according to claim 12,

wherein the postprocessor is configured to determine as the confidence information whether or not the at least one output value provided by the neural network indicates that the estimate of the loudness of the signal components of interest of the audio signal would be higher than a total loudness of the audio signal, and

wherein, if the at least one output value provided by the neural network indicates that the estimate of the loudness of the signal components of interest of the audio signal would be higher than a total loudness of the audio signal, the postprocessor is configured to modify the estimate of the loudness of the signal components of interest such that the loudness of the signal components of interest of the audio signal is equal to the total loudness of the audio signal.

14. The apparatus according to claim 12,

wherein the postprocessor is configured to determine as the confidence information whether or not the at least one output value provided by the neural network indicates that the estimate of the loudness of the signal components of interest of the audio signal would be higher than a total loudness of the audio signal, and

wherein, if the at least one output value provided by the neural network indicates that the estimate of the loudness of the signal components of interest of the audio signal would be higher than a total loudness of the audio signal, the postprocessor is configured to output the confidence information comprising an indication that the

estimate of the loudness of the signal components of interest of the audio signal is not reliable.

15. The apparatus according to one of claims 12 to 14,

wherein the postprocessor is configured to determine and to output the confidence information comprising a confidence value that indicates the degree of reliability of the estimate of the loudness of the signal components of interest of the audio signal conducted by the neural network, such that the confidence value depends on the estimate of the loudness of the signal components of interest of the audio signal and further depends on a loudness or an estimate of a loudness of the other signal components of the audio signal.

16. The apparatus according to claim 15,

wherein the confidence value depends on a difference between the estimate of the loudness of the signal components of interest of the audio signal and the loudness or the estimate of the loudness of the other signal components of the audio signal, or

wherein the confidence value depends on a ratio of the estimate of the loudness of the signal components of interest of the audio signal and the loudness or the estimate of the loudness of the other signal components of the audio signal.

17. The apparatus according to one of the preceding claims,

wherein the neural network has been trained using a plurality of data training items, and

wherein each of the plurality of data training items comprises one of a plurality of audio training signal portions and one or more reference loudness values.

18. The apparatus according to claim 17,

wherein the neural network has been trained depending on a loss function, and

wherein, to determine a return value of the loss function during training, the neural network is configured to determine one or more loudness value estimates of the

audio training signal portion for each of one or more data training items of the plurality of data training items, and

wherein the neural network has been trained depending on the loss function such that a return value of the loss function depends on the one or more loudness value estimates of the audio training signal portion and on the one or more reference loudness values of each of the one or more data training items.

19. The apparatus according to claim 18,

wherein one of the one or more reference loudness values of a data training item of the one or more data training items indicates a loudness of the signal components of interest of the audio training signal portion of the data training item, and wherein one of the one or more loudness value estimates of the data training item indicates an estimate of said loudness of the signal components of interest of the audio training signal portion of the data training item by the neural network; and/or

wherein one of the one or more reference loudness values of a data training item of the one or more data training items indicates a loudness of the other signal components of the audio training signal portion of the data training item, and wherein one of the one or more loudness value estimates of the data training item indicates an estimate of said loudness of the other signal components of the audio training signal portion of the data training item by the neural network; and/or

wherein one of the one or more reference loudness values of a data training item of the one or more data training items indicates a loudness of the entire audio training signal portion of the data training item, and wherein one of the one or more loudness value estimates of the data training item indicates an estimate of said loudness of the entire audio training signal portion of the data training item by the neural network; and/or

wherein one of the one or more reference loudness values of a data training item of the one or more data training items indicates a loudness of the audio training signal portion of the data training item when speech is present, and wherein one of the one or more loudness value estimates of the data training item indicates an estimate of said loudness of the audio training signal portion of the data training item by the neural network when speech is present, and/or

wherein one of the one or more reference loudness values of a data training item of the one or more data training items indicates a partial loudness of the signal components of interest the audio training signal portion of the data training item, and wherein one of the one or more loudness value estimates of the data training item indicates an estimate of said partial loudness of the signal components of interest of the audio training signal portion of the data training item by the neural network.

20. The apparatus according claim 18 or 19,

wherein the loss function is defined according to

$$Loss = \frac{1}{N}\sum_{i=1}^{N}(estimate_i - reference_i)^p$$

wherein $Loss$ indicates the return value of the Loss function,

wherein $estimate_i$ indicates one of the one or more loudness value estimates of an i-th data training item of the one or more data training items,

wherein $reference_i$ indicates one of the one or more reference loudness values of the i-th data training item of the one or more data training items, and

wherein p ≥ 1, and wherein N ≥1.

21. The apparatus according to one of claims 18 to 20,

wherein the neural network has been trained by iteratively adjusting the plurality of weights of the plurality of neural nodes of the neural network, and

wherein, in each iteration step of a plurality of iteration steps, the plurality of weights of the plurality of neural nodes of the neural network has been adjusted depending on one or more errors returned by the loss function in response to receiving the one or more data training items.

22. The apparatus according to one of claims 17 to 21,

wherein one of the one or more reference loudness values of one of the one or more data training items depends on one or more modified coefficients of the audio training signal portion of the data training item, and

wherein the one or more modified coefficients of the audio training signal portion of the data training item depends on one or more initial coefficients of the audio training signal portion of the data training item.

23. The apparatus according to claim 22,

wherein the one or more modified coefficients of the audio training signal portion of the data training item depend on an application of a filter on the one or more initial coefficients of said audio training signal portion, or

wherein the one or more modified coefficients of the audio training signal portion of the data training item depend on a spectral weighting of the one or more initial coefficients of the signal components of interest of said audio training signal portion.

24. The apparatus according to claim 23,

wherein the one or more modified coefficients indicate an squaring of each of one or more filtered coefficients which result from the application of the filter on the one or more initial coefficients, or

wherein the one or more modified coefficients indicate an squaring of each of one or more spectrally weighted coefficients which result from the spectral weighting of the one or more initial coefficients.

25. The apparatus according to claim 23 or 24,

wherein the filter depends on a psychoacoustic model, or

wherein the spectral weighting depends on the psychoacoustic model.

26. The apparatus according to one of claims 22 to 25,

wherein said one of the one or more reference loudness values depends on a sum or a weighted sum of at least two of the modified coefficients.

27. The apparatus according to claim 26,

wherein said one of the one or more reference loudness values depends on

$$L = a\ \left(\frac{1}{N}\sum_1^T x^2\right)^b$$

wherein x^2 indicates a square of a modified coefficient of the at least two of the modified coefficients,

wherein T is an integer indicating a number of the at least two of the modified coefficients,

wherein a and N are predefined numbers, and

wherein $0 < b < 1$.

28. The apparatus according to claim 26,

wherein said one of the one or more reference loudness values depends on

$$L = a\ log_b\left(\frac{1}{N}\sum_1^T x^2\right)$$

wherein x^2 indicates a square of a modified coefficient of the at least two of the modified coefficients,

wherein T is an integer indicating a number of the at least two of the modified coefficients,

wherein log indicates a logarithmic function being the compressive function, and

wherein a, b and N are predefined numbers.

29. The apparatus according to one of the preceding claims,

wherein the neural network comprises an input layer, two or more hidden layers, and an output layer, and

wherein the input layer comprises a plurality of input nodes, wherein each of the plurality of input nodes is configured to receive one of the plurality of input values, and

wherein each of the two or more hidden layers comprises one or more neural nodes, and

wherein the output layer comprises at least one output node, wherein the at least one output node is configured to output the at least one output value indicating the estimate of the loudness of the signal components of interest of the audio signal.

30. The apparatus according to claim 29,

wherein at least one layer of the two or more hidden layers is a convolutional layer.

31. The apparatus according to claim 30,

wherein the neural network is configured to employ a convolutional filter for the convolutional layer, which has a shape (x , y), with $x = y$ or with $x \neq y$, wherein $\max(x, y) \leq 10$.

32. The apparatus according to one of claims 29 to 31,

wherein at least one layer of the two or more hidden layers is a fully connected layer.

33. The apparatus according to one of claims 29 to 32,

wherein the hidden layers comprise at least one convolutional layer, at least one pooling layer, and at least one fully connected layer.

34. The apparatus according to one of claims 29 to 33,

wherein the apparatus is configured to employ linear activation in the output layer of the neural network.

35. The apparatus according to one of the preceding claims,

wherein the input interface is configured to receive a plurality of spectral samples of the audio signal as the plurality of input values, and

the neural network is configured to determine the estimate of the loudness of the signal components of interest of the audio signal depending on the plurality of power spectral samples of the audio signal.

36. The apparatus according to claim 35,

wherein the plurality of spectral samples are power spectral samples of at least 32 frequency bands.

37. The apparatus according to claim 35 or 36,

wherein the plurality of spectral samples of the audio signal represent the audio signal in a time-frequency domain.

38. The apparatus according to claim 37,

wherein the apparatus further comprises a transform module configured for transforming the audio signal from a time domain to the time-frequency domain to obtain the plurality of spectral samples of the audio signal.

39. The apparatus according to claim 38,

wherein the transform module is configured to transform segments of the audio signal of at least 100 ms length from the time domain to the time-frequency domain to obtain the plurality of spectral samples of the audio signal.

40. The apparatus according to one of claims 35 to 39,

wherein a first group of two or more of the plurality of spectral samples relate to a first group of frequency bands, which each exhibit a bandwidth that deviates by no more than 10% from a predefined first bandwidth, and

wherein a second group of two or more of the plurality of spectral samples relate to a second group of frequency bands, which each exhibit a higher center frequency than each frequency band of the first group of frequency bands, and which each

exhibit a bandwidth being higher than the bandwidth of each frequency band of the first group.

41. The apparatus according to claim 40,

wherein a third group of two or more of the plurality of spectral samples relate to a third group of frequency bands, which each exhibit a higher center frequency than each frequency band of the second group of frequency bands, which each exhibit a bandwidth being higher than the bandwidth of each frequency band of the second group, and

wherein the bandwidth of each frequency band of the third group deviates less from an equivalent rectangular bandwidth than the bandwidth of each frequency band of the second group.

42. A system for modifying an audio input signal to obtain an audio output signal, wherein the system comprises:

an apparatus according to one of the preceding claims for providing an estimate of a loudness of signal components of interest of the audio input signal, and

a signal processor configured to modify the audio input signal depending on the estimate of the loudness of the signal components of interest of the audio input signal to obtain the audio output signal.

43. A system according to claim 42,

wherein the signal components of interest of the audio signal are speech components of the audio signal, and

wherein the signal processor is configured to modify the audio input signal depending on the estimate of the loudness of the speech components of the audio input signal to obtain the audio output signal.

44. A system according to claim 43,

wherein the signal processor is configured to modify the audio input signal depending on the estimate of the loudness of the speech components of the audio

input signal and depending on an estimation of the loudness of the background components of the audio input signal to obtain the audio output signal.

45. A system according to claim 44,

wherein the apparatus for providing an estimate of a loudness of speech components of the audio input signal is an apparatus according to claim 11, and

wherein the signal processor is configured to modify a level of the audio input signal depending on the partial loudness of the speech components of the audio signal.

46. A method for providing an estimate of a loudness of signal components of interest of an audio signal, wherein the method comprises:

receiving a plurality of samples of the audio signal, and

estimating the loudness of the signal components of interest of the audio signal,

wherein a neural network receives as input values the plurality of samples of the audio signal or a plurality of derived values being derived from the plurality of samples of the audio signal, and

wherein the neural network determines at least one output value from the plurality of input values, such that the at least one output value indicates the estimate of the loudness of the signal components of interest of the audio signal, and

wherein the audio signal simultaneously comprises the signal components of interest and other signal components of the audio signal, and

wherein an influence of the other signal components on the estimate of the loudness of the signal components of interest is reduced or not present.

47. A computer program for implementing the method of claim 46 when being executed on a computer or signal processor.