

CLAIMS

1. A computer-implemented method of processing audio data, the method comprising:
 receiving input audio data (x) comprising a time-series of amplitude values;
 5 transforming the input audio data (x) into an input frequency band decomposition (X1) of the input audio data (x);
 transforming the input frequency band decomposition (X1) into a first latent representation (Z);
 processing the first latent representation (Z) by a first deep neural network to obtain a second latent representation (Z^Λ, Z1^Λ);
 transforming the second latent representation (Z^Λ, Z1^Λ) to obtain a discrete approximation (X3^Λ);
 10 element-wise multiplying the discrete approximation (X3^Λ) and a residual feature map (R, X5^Λ) to obtain a modified feature map, wherein the residual feature map (R, X5^Λ) is derived from the input frequency band decomposition (X1);
 processing a pre-shaped frequency band decomposition by a waveshaping unit to obtain a waveshaped frequency band decomposition (X1^Λ, X1.2^Λ), wherein the pre-shaped frequency band decomposition is derived from the input frequency band decomposition (X1), wherein the waveshaping unit comprises a second deep neural network;
 15 summing the waveshaped frequency band decomposition (X1^Λ, X1.2^Λ) and a modified frequency band decomposition (X2^Λ, X1.1^Λ) to obtain a summation output (X0^Λ), wherein the modified frequency band decomposition (X2^Λ, X1.1^Λ) is derived from the modified feature map; and
 transforming the summation output (X0^Λ) to obtain target audio data (y^Λ).
- 20 2. The method of claim 1, wherein transforming the input audio data (x) into the input frequency band decomposition (X1) comprises convolving the input audio data (x) with kernel matrix (W1).
3. The method of claim 2, wherein transforming the summation output (X0^Λ) to obtain the target audio data (y^Λ) comprises convolving the summation output (X0^Λ) with the transpose of the kernel matrix (W1T).
- 25 4. The method of any one of claims 1 to 3, wherein transforming the input frequency band decomposition (X1) into the first latent representation (Z) comprises locally-connected convolving the absolute value (|X1|) of the input frequency band decomposition (X1) with a weight matrix (W2) to obtain a feature map (X2); and max-pooling the feature map (X2) to obtain the first latent representation (Z).
- 30 5. The method of any one of claims 1 to 4, wherein the waveshaping unit further comprises a locally connected smooth adaptive activation function layer following the second deep neural network.
6. The method of claim 5, wherein the waveshaping unit further comprises a first squeeze-and-excitation layer following the locally connected smooth adaptive activation function layer.
- 35 7. The method of any one of claims 1 to 6, wherein at least one of the waveshaped frequency band decomposition (X1^Λ, X1.2^Λ) and the modified frequency band decomposition (X2^Λ, X1.1^Λ) is scaled by a gain factor (se, se1, se2) before summing to produce the summation output (X0^Λ).
- 40 8. The method of any one of claims 2 to 7, wherein each of kernel matrix (W1) and the weight matrix (W2) comprises

fewer than 128 filters, optionally fewer than 32 filters, optionally fewer than 8 filters.

9. The method of any one of claims 1 to 8, wherein the second deep neural network comprises first to fourth dense layers optionally respectively comprising 32, 16, 16 and 32 hidden units, optionally wherein each of the first to third dense layers of the second deep neural network is followed by a tanh function.
10. The method of any one of claims 6 to 9, wherein, in the waveshaping unit, the first squeeze-and-excitation layer comprises an absolute value layer preceding a global average pooling operation.
11. The method of any one of claims 1 to 10, further comprising:
 passing on the input frequency band decomposition (X1) as the residual feature map (R);
 passing on the modified feature map as the pre-shaped frequency band decomposition; and
 passing on the modified feature map as the modified frequency band decomposition (X2^Δ, X1.1^Δ).
12. The method of claim 11, wherein the first deep neural network comprises a plurality of bidirectional long short-term memory layers, optionally followed by a smooth adaptive activation function layer.
13. The method of claim 12, wherein the plurality of bidirectional long short-term memory layers comprises first, second and third bidirectional long short-term memory layers, optionally comprising 64, 32 and 16 units respectively.
14. The method of claims 12 or 13, wherein the plurality of bidirectional long short-term memory layers is followed by a plurality of smooth adaptive activation function layers, each optionally being composed of 25 intervals between -1 to +1.
15. The method of claim 12, wherein the first deep neural network comprises a feedforward WaveNet comprising a plurality of layers, optionally wherein the final layer of the WaveNet is a fully-connected layer.
16. The method of any one of claims 1 to 10,
 wherein the first deep neural network comprises a plurality of shared bidirectional long short-term memory layers, followed by, in parallel, first and second independent bidirectional long short-term memory layers;
 wherein the second latent representation (Z1^Δ) is derived from the output of the first independent bidirectional long short-term memory layer;
 wherein, in the waveshaping unit, the first squeeze-and-excitation layer further comprises a long short-term memory layer;
 wherein the method further comprises:
 passing on the input frequency band decomposition (X1) as the pre-shaped frequency band decomposition;
 processing the first latent representation (Z) using the second independent bidirectional long short-term memory layer to obtain a third latent representation (Z2^Δ);
 processing the third latent representation (Z2^Δ) using a sparse finite impulse response layer to obtain a fourth latent representation (Z3^Δ);
 convolving the frequency band representation (X1) with the fourth latent representation (Z3^Δ) to obtain said residual feature map (X5^Δ); and

processing the modified feature map by a second squeeze-and-excitation layer comprising a long short-term memory layer to obtain said modified frequency band decomposition ($X2^\wedge$, $X1.1^\wedge$).

17. The method of claim 16, wherein the plurality of shared bidirectional long short-term memory layers comprises first and second shared bidirectional long short-term memory layers optionally comprising 64 and 32 units respectively, optionally wherein each of the first and second shared bidirectional long short-term memory layers has a tanh activation function.
18. The method of claims 16 or 17, wherein each of the first and second independent bidirectional long short-term memory layers comprises 16 units, optionally wherein each of the first and second independent bidirectional long short-term memory layers comprises a locally connected smooth adaptive activation function.
19. The method of any one of claims 16 to 18, wherein the sparse finite impulse response layer comprises:
first and second independent dense layers taking the third latent representation ($Z2^\wedge$) as input; and
a sparse tensor taking the respective output of the first and second independent dense layers as inputs, the output of the sparse tensor being the fourth latent representation ($Z3^\wedge$);
optionally wherein the first and second independent dense layers comprise respectively a tanh function and a sigmoid function.
20. The method of any one of claims 1 to 14 and 16 to 19, wherein all the convolutions are along the time dimension and have a stride of unit value.
21. The method of any one of claims 1 to 14 and 16 to 20, wherein at least one of the deep neural networks is trained in dependence on data representing one or more audio effect selected from a group comprising: tube amplifier, distortion, speaker-amplifier, ladder filter, power amplifier, equalisation, equalisation-and-distortion, compressor, ring modulator, phaser, modulation based on operational transconductance amplifier, flanger with bucket brigade delay, modulation based with bucket brigade delay, Leslie speaker horn, Leslie speaker horn-and-woofer, flanger-and-chorus, modulation based, modulation based-and-compressor, plate-and-spring reverberation, echo, feedback delay, slapback delay, tape-based delay, noise-driven stochastic effects, dynamic equalisation based on input signal level, audio morphing, timbre transformations, phase vocoder, time stretching, pitch shifting, time shuffling, granulation, 3D loudspeaker setup modelling, and room acoustics.
22. A computer program comprising instructions which, when the program is executed by a computer, cause the computer to carry out the method of claims 1 to 21.
23. A computer-readable storage medium comprising the computer program of claim 22.
24. An audio data processing device comprising a processor configured to perform the method of claims 1 to 21.