GRUs may be built one over the other and this time the output of one GRU is conveyed to the input of the next GRU. Another alternative could be to also connect the states between the cascaded recurrent units with mechanisms such an attention.

The relationships may be governed, for example, by formulas such as at least one of the following:

$$z_t = \sigma\left(W_z \cdot (h_{t-1}, x_t)\right)$$
$$r_t = \sigma\left(W_r \cdot (h_{t-1}, x_t)\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot (r_t * h_{t-1}, x_t)\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

where:

t refers to the time instant/step, and in case of unrolled GRUs, refers also to the particular module in the unrolled structure (e.g. t=0 is the first module, first time instant, t=1 the second, and so on);

x_t refers to the input vector of the recurrent module at instant t (e.g. to the frame at the time t, e.g. with or without the samples taken from the (e.g. immediately) preceding frame and/or with or without the samples taken from the (e.g. immediately) preceding frame);

h_t refers to the state and output at instant t of the recurrent unit, which will be inherited by the $(t+1)^{th}$ feedforward module in the unrolled case (with reference to Fig. 11, h_t is reintroduced in feedback as h_{t-1}, see below; with reference to Fig. 12, h_t is provided to the immediately subsequent feedforward module);

h_{t-1} refers to the state and output at time step t-1, which is the input of the unit at instant t. In case of unrolled GRU (Fig. 12), h_{t-1} is an input of tth feedforward module (i.e., either the output of the immediately preceding recurrent module, or the input of the GRU) (if the tth module is the first module, then h_{t-1} will be a default value);

$\widetilde{h_t}$ refers to a candidate state and/or output of the recurrent module;

z_t refers to an update gate vector;

r_t refers to a reset gate vector;

We report the computational complexity estimates in the following table.

Model	Encoder Complexity	Decoder Complexity
SSMGAN	0.05 GMACs	4.56 GMACs
SoundStream	5 GMACs	5 GMACs
NESC	0.5 GMACs	7 GMACs

Table 1: *Complexity Estimation*

Our implementation runs faster than real time on a single thread of an Intel(R) Core(TM) i7-6700 CPU at 3.40Ghz

3.3. Qualitative statistical analysis of the latent

We provide a qualitative analysis of the distribution of the latent in order to give a better understanding of its behaviour in practice. The quantized latent frames are embedded in a space of dimension 256 hence in order to plot their distribution we use their t-SNE projections. For each experiment we first encode 10 s of audio with different recording conditions and we label each frame depending on a priori information regarding its acoustic and linguistic characteristics. Each subplot represent a different set of audio randomly selected from the LibriTTS, VCTK and NTT Datasets. Afterwards we look for clusterings in the low dimensional projections. Notice that the model is not trained with any clustering objective, hence any such behaviour shown at inference time is an emergent aspect of the training set up.

We test both speaker characteristics, such as language and gender, and acoustic aspects like voicing and noisiness. In our first experiment (Fig. 2a) we test voicing information using a VAD algorithm to label each frame automatically. We notice a clear clustering of voiced, unvoiced and silent frames with the boundary consisting of transition frames. We similarly label voiced frames based on their quantized pitch values, but this shows no significant clustering behaviour. We do not show the picture because of lack of space.

Codec	POLQA	STOI	ViSQOL v3
OPUS 6 kbps	1.681	0.480	2.273
EVS 5.9 kbps	**3.308**	**0.553**	**3.036**
SSMGAN 1.6 kbps	2.213	0.536	2.505
NESC 1 kbps	1.534	0.612	2.109
NESC 2 kbps	2.382	0.641	2.615
NESC 3 kbps	**2.548**	**0.643**	**2.841**

Table 2: Average objective scores for neural decoders on the In-formalSet. For all metrics higher scores are better. Confidence intervals are negligible for POLQA and ViSQOL v3, while for STOI they are smaller than 0.025

Codec	POLQA	STOI	ViSQOL v3
OPUS 6 kbps	1.833	0.613	2.357
EVS 5.9 kbps	**3.486**	**0.736**	**3.071**
SSMGAN 1.6 kbps	2.267	0.647	2.476
NESC 1 kbps	1.659	0.745	2.074
NESC 2 kbps	2.492	0.802	2.595
NESC 3kbps	**2.707**	**0.817**	**2.822**

3.5. Subjective Evaluation

We test the model only on challenging unseen conditions in order to assess its robustness. For this we select a test set of speech samples from the NTT Dataset comprising unseen speakers, languages and recording conditions. In the test set "m" stands for male, "f" for female, "ar" for Arabic, "en" for English, "fr" for French, "ge" for German, "ko" for Korean, and "th" for Thai.

We also test the model on noisy speech. For this we select the same speech samples as for the clean speech test and apply a similar augmentation policy as in Section 3.1. We add ambient noise samples (e.g. airport noises, typing noises, ...) at SNR between 10 dB and 30 dB and then convolve with room impulse responses (RIR) coming from small, medium and big sized rectangular rooms. More precisely, "ar/f", "en/f", "fr/m", "ko/m", and "th/f" are convolved with RIRs from small rooms,

[19] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural discrete representation learning,” in Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 6309–6318.

[20] H. Zen, V. Dang, R. Clark, Y. Zhang, R. Weiss, Y. Jia, Z. Chen, and Y. Wu, “Libritts: A corpus derived from librispeech for text-to-speech,” arXiv preprint arXiv:1904.02882, 2019.

[21] C. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, “Icassp 2021 deep noise suppression challenge,” in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 6623–6627.

[22] D. Povey, “OpenSLR 28 Dataset,” https://www.openslr.org/28//.

[23] R. Yamamoto, E. Song, and J. Kim, “Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram,” in ICASSP 2020, IEEE International Conference on Acoustics, Speech and Signal Processing, 2020, pp. 6199–6203.

[24] K. Kumar, R. Kumar, de T. Boissiere, L. Gestin et al., “MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis,” in Advances in NeurIPS 32, 2019, pp. 14 910–14 921.

[25] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” ICLR, 2015.

[29] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "ViSQOL v3: An open source production ready objective speech and audio metric," in *2020 twelfth international conference on quality of multimedia experience (QoMEX)*. IEEE, 2020, pp. 1- 6.

[30] J. Beerends, C. Schmidmer, J. Berger, M, Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual Objective Listening Quality Assessment (POLQA), *the* third generation ITU-T standard for end-to-end speech quality measurement part I - temporal alignment," *Journal of the audio engineering society,* vol. 61, no. 6 , pp. 1116-184, june. 2013.

[31] C. H. Taal, R. C. Hendriks. R, Heusdens, and J. Jensen, "Algorithm for intelligibility prediction of time frequency weighted noisy speech," *IEEE Trans Audio Speech Lang. Process., vol,* pp.2125- 2 136. 201 I.

Further characterization of the figures

Figures 1b and 8: Neural End-2-End Speech Codec high level architecture.

As shown in examples above, there is disclosed a method for training the audio signal generator [e.g. decoder], may comprise a training session including generating a multiplicity of bitstreams with candidate indexes associated with known frames representing known audio signals, the training session including a decoding of the bitstreams and an evaluation of audio signals generated by the decoding in respect to the known audio signals, so as to adapt associations of indexes of the at least one codebook with the frames of the encoded bitstreams [e.g. by minimizing the difference between the generated audio signal and the known audio signals] [e.g. using a GAN].

As shown in examples above, there is disclosed a method for training an audio signal generator [e.g. decoder] as above, may comprise a training session including generating a multiplicity of bitstreams with candidate indexes associated with known frames representing known audio signals, the training session including providing to the audio signal generator bitstreams non-provided by the encoder, so as to obtain the indexes to be used [e.g. obtain the codebook] by optimizing a loss function.

As shown in examples above, there is disclosed a method for training an audio signal generator [e.g. decoder] as above, may comprise a training session including generating multiple output audio signal representations of known input audio signals, the training session including an evaluation of the multiple output audio signal representations [e.g. bitstreams] in respect to the known input audio signals and/or minimizing a loss function, so as to adapt parameters of at least one learnable layer(s) optimizing a loss function.

As shown in examples above, there is disclosed a method for training an audio signal representation generator (or encoder) as above, may comprise a training session including receiving a multiplicity of bitstreams with indexes associated with known frames representing known audio signals, the training session including an evaluation of the generated audio signals in respect to the known audio signals, so as to adapt associations of indexes of the at least one codebook with the frames of the encoded bitstreams and/or optimizing a loss function [e.g. by minimizing the

difference between the generated audio signal and the known audio signals] [e.g. using a GAN].

As shown in examples above, there is disclosed a method for training an audio signal representation generator (or encoder) as above together with an audio signal generator [e.g. decoder] e.g. as above, may comprise:

providing a plurality of audio signals (1) to the audio signal representation generator, so as to obtain audio signal representations and/or bitstreams (3) and, at the audio signal generator (10), generating the output signals (16) from the audio signal representations and/or bitstreams (3);

providing, to the audio signal generator (10), a plurality of audio signal representations and/or bitstreams (3) which not generated by the audio signal representation generator (20), and, at the audio signal generator (10), generating the output signals (16) from the audio signal representations and/or bitstreams (3);

evaluating a loss function associated to the output signals (16) from the audio signal representations and/or bitstreams (3) vs the output signals (16) from the audio signal representations and/or bitstreams (3), so as to obtain the parameters of the learnable layers and blocks of the audio signal generator (10) and of the audio signal representation generator by minimizing the loss function.

As shown in examples above, there is disclosed a method for generating an audio signal (16) from a bitstream (3), the bitstream (3) representing the audio signal (16), the audio signal being subdivided in a sequence of frames, the method may comprise:

providing, for a given frame, first data (15) derived from an input signal (14)[e.g. from an external or internal source or from the bitstream (3)], [wherein the first data (15) may have one single channel or multiple channels];

though a first processing block (40, 50, 50a-50h), receiving [e.g. for the given frame] the first data (15) and outputting first output data (69) in the given frame, [wherein the first output data (69) may comprise a one single channel or a plurality of channels (47)],

[e.g. the method also comprising through a second processing block (45), e.g. for the given frame, receiving, as second data, the first output data (69) or data derived from the first output data (69),]

wherein the first processing block (50) comprises:

Claims

1. An audio signal representation generator for generating an output audio signal representation from an input audio signal including a sequence of input audio signal frames, each input audio signal frame including a sequence of input audio signal samples, the audio signal representation generator comprising:

a format definer configured to define a first multi-dimensional audio signal representation of the input audio signal, the first multi-dimensional audio signal representation of the input audio signal including at least:

a first dimension, so that a plurality of mutually subsequent frames is ordered according to the first dimension; and

a second dimension so that a plurality of samples of at least one frame are ordered according to the second dimension,

at least one learnable layer configured to process the first multidimensional audio signal representation of the input audio signal, or processed version of the first multi-dimensional audio signal representation, to generate the output audio signal representation of the input audio signal,

wherein the format definer is configured to insert, along the second dimension of the first multi-dimensional audio signal representation of the input audio signal, additional input audio signal samples of one or more additional frames immediately successive to or immediately preceding the given frame.

2. The audio signal representation generator of claim 1, wherein the format definer is configured to insert, along the second dimension of the first multidimensional audio signal representation of the input audio signal, input audio signal samples of each given frame.

3. The audio signal representation generator of claims 1 or 2, wherein the at least one learnable layer includes at least one recurrent learnable layer.

4. The audio signal representation generator of claim 3, wherein the at least one recurrent learnable layer includes a gated recurrent unit, GRU.

5. The audio signal representation generator of claim 3 or 4, wherein the at least one recurrent learnable layer is operated along the first dimension.

6. The audio signal representation generator of any of the preceding claims, further comprising at least one first convolutional learnable layer between the format definer and the at least one recurrent learnable layer.

7. The audio signal representation generator of claim 6, wherein in the at least one first convolutional learnable layer the kernel is slid along the second direction of the first multi-dimensional audio signal representation of the input audio signal.

8. The audio signal representation generator of any of the preceding claims, further comprising at least one convolutional learnable layer downstream to the at least one recurrent learnable layer.

9. The audio signal representation generator of claim 8, wherein in the at least one convolutional learnable layer the kernel is slid along the second direction of the first multi-dimensional audio signal representation of the input audio signal.

10. The audio signal representation generator of any of the preceding claims, wherein at least one or more of the at least one learnable layer is a residual learnable layer.

11. The audio signal representation generator of claim 10, wherein at least one learnable layer is a residual learnable layer, a main portion of the first multidimensional audio signal representation of the input audio signal bypassing the at least one learnable layer, and/or the at least one learnable layer is applied to at least a residual portion of the first bidimensional audio signal representation of the input audio signal.

12. The audio signal representation generator of any of claims 3-11, wherein the recurrent learnable layer operates along a series of time steps each having at least one state, in such a way that each time step is conditioned by the output and/or state of the preceding time step.

13. The audio signal representation generator of claim 12, wherein the step and/or output of each step is recursively provided to a subsequent time step.

14. The audio signal representation generator of claim 12 or 13, comprising a plurality of feedforward modules, each providing the state and/or output to the subsequent module.

15. The audio signal representation generator of any of claims 3-14, wherein the recurrent learnable layer generates the output for a given time instant by keeping into account the output and/or a state of a preceding time instant, wherein the relevance of the output and/or state of a preceding time instant is obtained training.

16. The audio signal representation generator of any of the preceding claims, the format definer begin configured to order mutually subsequent samples, one after the other one according to the second dimension.

17. An encoder comprising an audio signal representation generator according to any of the preceding claims and a quantizer to encode a bitstream from the output audio signal representation.

18. The encoder of claim 17, wherein the quantizer is a learnable quantizer configured to associate, to each frame of the first multi-dimensional audio signal representation of the input audio signal, or a processed version of the first multi-dimensional audio signal representation, indexes of at least one codebook, so as to generate the bitstream.

19. The encoder of claim 18, wherein the learnable quantizer uses the at least one codebook associating indexes i_z, i_r, i_q, with the index i_z representing a code z approximating E(x) and being taken from the codebook z_e, the index i_r representing a code r approximating E(x)-z and being taken from the codebook r_e, and the index i_q representing a code q approximating E(x)-z-r and being taken from the codebook q_e to be encoded in the bitstream.

20. The encoder of any of claims 18-19, wherein the at least one codebook includes at least one base codebook associating, to indexes to be encoded in the bitstream, multidimensional tensors of the first multi-dimensional audio signal representation of the input audio signal.

21. The encoder of any of claims 18-20, wherein the at least one codebook includes at least one residual codebook associating, to indexes to be encoded in the bitstream, multidimensional tensors of the first multi-dimensional audio signal representation of the input audio signal.

22. The encoder of any of claims 18-21, wherein there are defined a multiplicity of residual codebooks, so that:

a second residual codebook associates, to indexes to be encoded in the audio signal representation, multidimensional tensors representing second residual portions of the first multi-dimensional audio signal representation of the input audio signal,

a first residual codebook associates, to indexes to be encoded in the audio signal representation, multidimensional tensors representing first residual portions of frames of the first multi-dimensional audio signal representation,

wherein the second residual portions of frames are residual with respect to the first residual portions of frames.

23. The encoder of any of claims 17-22, configured to signal, in the bitstream, whether indexes associated to residual frames are encoded or not.

24. The encoder of any of claims 18-23, wherein at least one codebook is a fixed-length codebook.

25. The encoder of any of claims 17-24, further comprising at least one further learnable block downstream to the at least one learnable block to generate, from the fourth multi-dimensional audio signal representation or another version of the input audio signal, a fifth audio signal representation of the input audio signal with multiple samples for each frame.

26. The encoder of claim 25, wherein the at least one further learnable block downstream to the at least one learnable block includes:

at least one residual learnable layer.

27. The encoder of claim 25 or 26, wherein the at least one further learnable block downstream to the at least one learnable block includes:

at least one convolutional learnable layer.

28. The encoder of any of claims 25-27, wherein the at least one further learnable block downstream to the at least one learnable block includes:

at least one learnable layer activated by an activation function (e.g. ReLu or Leaky ReLu).

29. A method for generating an output audio signal representation from an input audio signal including a sequence of input audio signal frames, each input audio signal frame including a sequence of input audio signal samples, the audio signal representation generator comprising:

defining a first multi-dimensional audio signal representation of the input audio signal;

through a first learnable layer, generating a second multi-dimensional audio signal representation of the input audio signal by sliding along a second direction of the first multi-dimensional audio signal representation of the input audio signal;

through a second learnable layer which is a recurrent learnable layer generating a third multi-dimensional audio signal representation of the input audio signal by operating along a first direction of the second multi-dimensional audio signal representation of the input audio signal;

through a third learnable layer which is a convolutional learnable layer generating a fourth multi-dimensional audio signal representation of the input audio signal by sliding along the second direction of the first multi-dimensional audio signal representation of the input audio signal,

so as to obtain the output audio signal representation from the fourth multi-dimensional audio signal representation of the input audio signal,

wherein the defining includes inserting, along the second dimension of the first multi-dimensional audio signal representation of the input audio signal, additional input audio signal samples of one or more additional frames immediately successive to or immediately preceding the given frame.

30. A non-transitable storage unit storing instruction which, when executed by a processor, cause the processor to perform a method according to claim 29.