Adaptive VR Test in Music Harmony Based on Conditional Spiking GAN

Anna Shvets
Fablab by Inetum
157 Boulevard MacDonald
75019 Paris, France
anna.shvets@inetum.com

Samer Darkazanli
iMSA
Rue Clos Maury 82000 Montauban,
France
darkazanli.samer@imsa.msa.fr

This article proposes an adaptive VR test for the knowledge level control in music harmony. The core functioning relies on conditional semantic music generation strategy, using spiking conditional GAN architecture. The novel method of semantic music information encoding based on the system of graphs in music harmony, allowed two-dimensional data representation of harmonic sequences, which made possible considerable data augmentation and a transition to the specifics of training inherent to the visual domain. To our best knowledge, this is the first attempt of conditional spiking GAN implementation along with the application of the spiking neural networks in a domain of semantic music generation.

1. INTRODUCTION

The search for sustainable and scalable calculation platforms led to the emergence of neuromorphic devices along with neuromorphic computing methods. The biologically plausible spiking neural networks (SNNs), considered as the 3rd generation of artificial neural networks, represent a viable alternative to the 2nd generation artificial neural networks (ANNs), as their structure and functioning are in a direct correspondence with the goals and technical possibilities of the neuromorphic devices. There are three strategies for spiking neural networks training, which comprise a posteriori conversion of a trained ANN into SNN (Massa et al. 2020, Diehl et al. 2016), designing and training SNNs in a spiking domain on conventional computing platforms, such as GPU or TPU (Rathi & Roy 2020) and training SNNs directly on low power devices (Akbarzadeh-Sherbaf, Safari & Vahabie 2020). In this article we exploit the second strategy, designing a conditional spiking GAN, training it on a GPU platform and integrating a trained model into an adaptive VR test scenario.

2. STATE OF THE ART

2.1 Spiking neural networks

2.1.1 Encoding strategies

The ways of the information encoding strategies, since the appearance of the first scientific model of spiking neural networks (SNNs) in 1952, comprise binary coding, rate coding, latency coding and fully temporal codes. The information encoding influences the learning methods taxonomy discussed below.

2.1.2 Learning methods

The application of the training strategy for SNNs depends on the nature of the problem and can be solved with unsupervised, supervised and reinforcement learning. The set of unsupervised learning methods comprise spike-timing-dependent plasticity – STDP rule (Caporale & Dan 2008), Growing Spiking Neural Networks (Hazan et al. 2008), Hebbian learning rule (Hebb 1949) with two derivatives - Artola, Bröcher, Singer – ABS rule (Artola & Singer 1993) and Bienenstock, Cooper, Munro – BCM rule (Bienenstock, Cooper & Munro 1982). Among them, the most used method is the STDP, which implies that the weight (synaptic efficacy) connecting pre-synaptic and post-synaptic neurons is altered based on their relative spike times, thus the weight adjustment is made using local information in terms of synapse and time.

The supervised learning methods comprise SpikeProp (Bohté, Kok & Poutré 2000), Remote Supervised Method – ReSuMe (Ponulak & Kasinski 2006), FreqProp (Bogacz, Brown & Giraud-Carrier 2000) and Local Error-Driven Associative
Biologically Realistic Algorithm – LEABRA (O’Reilly 1996). More recently, a latency-based backpropagation for static stimuli – S4NN with surrogate gradient learning (Kheradpisheh & Masquelier 2020), binarized spiking neural networks with temporal coding and learning - BS4NN (Kheradpisheh, Mirsadeghi, & Masquelier 2021) and rectified linear postsynaptic potential function (Zhang et al. 2021) have proposed a viable alternative to the existing methods, by adapting backpropagation algorithm to the SNN training specifics.

Finally, the reinforcement learning makes usage of the spiking actor-critic method (Potjans, Morrison & Diesmann 2009) and through reward-modulated STDP (Florian 2007).

2.1.3 SNN neuron architecture

The mathematical formalism of the biological SNN neurons can be divided into two groups: conductance-based models and threshold models. Conductance-based models, such as Hodgkin-Huxley model (Hodgkin & Huxley 1952), FitzHugh–Nagumo model (Fitzugh 1961), Morris–Lecar model (Morris & Lecar 1981), Hindmarsh–Rose model (Hindmarsh & Rose 1984), Izhikevich model (Izhikevich 2003) or Cable theory (Tuckwell 1988), describe the initiation and propagation of the action potentials in neurons, while threshold models, such as perfect (non-leaky) integrate-and-fire, leaky integrate-and-fire (Delorme et al. 1999) or adaptive exponential integrate-and-fire (Brette & Gerstner 2005), generate an impulse while a certain threshold is reached. Recent research is mostly exploits the threshold models, as per the simplicity of their calculation.

2.1.4 Network architectures

The integration of the SNN neurons has been tested with classical feedforward (dense) neural networks (She 2020), recurrent neural networks (Demirag et al. 2021, Kim & Sejnowski 2019), convolutional neural networks (Guan & Mo 2020) and belief neural networks (O’Connor et al. 2013). The SNN layers were also applied within generative adversarial network architecture and are discussed below.

2.2 Conditional music generation and generative potential of SNNs

2.2.1 Conditional music generation

The attempts to control the generated samples may be divided into three groups – conditional, controllable and constraint generation. The conditional generation takes one element to generate another (Liu & Yang 2018, Yu & Canales 2021), while controllable generation uses the input features change to manipulate some aspects of the output generation (Wang et al. 2020, Tan & Herremans 2020). Finally, the constraint generation exploits the template-based approach to influence a shape of the output result (Lattner, Grachten & Widmer 2016). The controllability research mainly focuses on the features disentanglement, proposing systematic studies (Pati & Lerch 2021) and datasets (Pati, Gururani & Lerch 2020), designed to foster further experiments in the field. The existing resources, however, gather monophonic music examples only and are not suitable for harmonic sequences studies.

The research in conditional music generation presents a plethora of generative architectures: LSTM, Transformer (Makris, Agres & Herremans 2021), GAN (Liu & Yang 2018, Shvets & Darkazanli 2021), hybrid versions, such as LSTM-GAN (Yu & Canales 2021) or GAN with an inception model (Li & Sung 2021), the latter architecture makes use of the convolutional layers, followed by the time distribution layer that considers sequential data, forcing the convolutional layers consider the time relationship in a manner similar to RNN layers do. The above-mentioned approaches make use of the time information encoding and this is an important point of attachment with the spiking neural networks, which are intrinsically sensitive to the temporal characteristics of information transmission (Tavanaei et al. 2019).

2.2.2 Spiking GANs

The current state of art in the application of spiking layers inside a GAN architecture accounts only three experiments – Spike-GAN (Molano-Mazon et al. 2018), Spiking-GAN (Kotariya & Ganguly 2021) and SpikeGAN (Rosenfeld, Simeone & Rajendran 2021).

The purpose of the Spike-GAN consisted in synthesizing of neural responses that approximate the statistics of the realistic neural activity, being trained on a real dataset recorded from the salamander retina (8192 samples) in a form of one-dimensional matrices of size $N \times T$, where $N$ represents the number of neurons and $T$ stands for the number of time bins during which the spikes occurred. Thus, the architecture of the Spike-GAN consisted of the discriminator with 1D convolutional layers (256 and 512 features, respectively), followed by a linear layer, along with a mirrored generator architecture sampling from a 128-dimension uniform distribution. The LeakyReLU activation function was used consistently through the network.

Spiking-GAN, instead of approximating neural activity and retrieving its heat maps, applies spiking learning strategy to the generation of examples inherent to a visual domain, using standard MNIST dataset (60 000 training examples). The experiment is based on Time-to-first-spike (TTFS)
temporal coding and the least-squares loss applied in the temporal domain. The Spiking-GAN architecture makes use of the dense layers: 2-layer fully connected network for generator (100-400-784) and discriminator (784-400-2). The latter takes a flattened spike train, encoded with TTFS coding, as an input. The output neurons of generator and discriminator are using tangent and sigmoid activation functions respectively. The rest of the activations used through the network are ReLU.

Finally, SpikeGAN aims matching the distribution of the SNN outputs with a target distribution, regardless of the data nature, being trained consistently on handwritten digits, simulated spike-domain handwritten digits and synthetic temporal data. SpikeGAN exploits a hybrid architecture with a spiking generator and conventional ANN discriminator.

None of the above-mentioned experiments, however, proposed a conditional spiking GAN architecture applied in a field of semantic music generation.

3. ADAPTIVE TEST DESCRIPTION

3.1 Methodological context

This proposal is based on a new system of representation, which lifts a cognitive load regarding the understanding of harmonic logic and chord structure, facilitating the assimilation of knowledge in musical harmony and the formation of audio skills. The method itself is based on a graph theory (Minsky 1974) and uses of the original mapping of colour to the functions of music harmony and to the chord structure. The effectiveness of this representation methodology has been proven efficient in a multi-step educational experiment on hybrid learning, showing a substantial increase of the quality of knowledge with the system application (Pistone & Shvets 2014, Jemielnik & Shvets 2015, Shvets 2019). The method was used to build graphic interface of the award-winning mobile and VR applications (Shvets 2016, Shvets & Darkazanli 2020).

3.2 Functioning scheme

The technical issues of implementing an adaptive test in music harmony previously was linked to the absence of reliable techniques for the semantic music content generation conditioned by the user's knowledge. In the present proposal, this problem is solved with the generative neural network with the conditional spiking GAN architecture, which takes the first chord as an input and generates a sequence of five chords. A conditional GAN with the similar functioning was proposed in 2021 (Shvets & Darkazanli 2021) and consisted of convolutional 1D layers, being integrated into a practice room of the VR applying “Graphs in music harmony”. In the present model, we add spiking layers, which improve learning spatiotemporal relationships and therefore memorizing the position of the chord in the sequence. The integration of this technology into the adaptive testing process consists of the following steps:

(i) Analysis of user data to define the set of chords that the user has already learned, using the internal storage of the VR application.
(ii) Generation of the harmonic sequence by the GAN model trained and deployed in the cloud, conditioned by the analysis of the learned chords. The data exchange protocol uses JSON format to send the user data and receive a generated sequence.
(iii) The sequence received from the model in a JSON format is mapped to a 3D representation of chords in a VR space, using the internal mapping rules. The object of the test (a chord) is replaced by a question mark symbol on the staff modelled in a VR space and the audio of the harmonic sequence, constructed from the VR application’s internal sound library, is played to the user.
(iv) The user chooses a chord from the graph, which matches the chord hidden by a question mark, played to the user previously. The VR application compares user input with the information received from the model.
(v) Depending on the correctness of the response, the VR test shows a chosen chord within the graph sequence (Fig. 2) and sends a new request to the generative model for generation of a new harmonic sequence. The generation might be requested either with another test object (a hidden chord), among those which should be learned in a lesson (in case of the correct answer), or for the same test object - the chord which was not recognised by the user (in case of the incorrect answer).
(vi) The transition to the end of the session is made after several repetitions of the step 3 (this number is defined by the number of chords to be learned per lesson), if all the answers are correct, or after the same number of repetitions, multiplied by the number of incorrect answers from the user.
Adaptive VR Test in Music Harmony Based on Conditional Spiking GAN
Anna Shvets & Samer Darkazanli

Figure 1: Scheme of functioning

Figure 2: Screenshot of the answer

3.3 Individual tones interactivity
To increase the immersion effect, the interactivity with the individual tones of the chord has been introduced. This feature allows the feedback reception (sound and vibration) after touching different tones of the chord. The feature became technically possible with the Oculus Quest 2 controller. This will allow the increased tangibility of chords and the possibility of audio separation of different chord tones, explaining the chord phonism.

4. CONDITIONAL SPIKING GAN
4.1 Data encoding
Since the nature of the model purpose, which consists in inclusion of chords learned during the lesson, each harmonic progression must contain the passing progression in a single tonality – C major. This very rigid limitation induced the necessity of manual data crafting and a search for the effective augmentation techniques. In this context, we propose a new encoding strategy, which converts one-dimensional textual data (analytical representation of a harmonic sequence) into two-dimensional spike trains. The base for the dimensionality shift is the conversion of the harmonic sequences to their matrix representation within the system of graphs. In order to perform a dimensionality transition, the following steps should be made:

(i) Mapping of each chord of the harmonic sequence in its analytical semantic representation to the numerical representation – respective pointer indexes;
(ii) Finding a corresponding chord within the system of graphs and replacing it with the mapped numerical value;
(iii) Applying a varying normalization term of a small value (1 2\textsuperscript{nd}) to augment the data.

To illustrate the described transition process, let us take a sequence of three chords II\textsubscript{7}-VI\textsubscript{46}-II\textsubscript{56} and present it as a graph (Fig. 3a), then replace a semantic chord designation with numerical values – indexes we chose to represent the chords (Fig. 3b): we thus receive a 3x3 matrix. If we repeat the described procedure, but taking the dimensionality
of the whole system, we receive a 28x28 matrix, which becomes a feature map to the neural network (the 49 chords of the system of graphs are giving precisely 27x21 matrix, however we added one column and seven rows of a padding filled with zeros to facilitate computing of the 2D convolutional operations within the neural network).

\[
\begin{array}{c|c|c}
& II_7 & \text{III}_{35} \\
V_{I_6} & 0 & 1 \\
II_{35} & 2 & 0 \\
\end{array}
\]

\[a) \quad b) \]

**Figure 3:** Harmonic sequence in a graph representation (a) and in a matrix representation (b).

The normalization variation allowed to augment the data from 216 manually crafted harmonic sequences of 5 chords each to 2160 matrices of training data. The sparsity of data points in each of the matrices is coherent with the advantage of spiking models, which is temporal sparsity – the fig. 4 shows a visualisation of such sparsity within ten matrices.

\[\text{Figure 4: Ten harmonic sequences visualised as a features map in ten matrices.}\]

### 4.2 Network architecture

The network was implemented with the use of Pytorch and PyTorch-Spiking library developed by Nengo.AI group. The latter framework allows transforming standard Pytorch activation functions into spiking layers with the possibility to use the spiking activations on the forward pass, using user defined time steps, and the non-spiking activation function on the backward pass, which overcomes the non-differentiability problem of spiking neural networks.

The generator model consists of a tripled stack of convolutional transposed 2D layer and a spiking layer (a ReLU activation function wrapper), the last stack contains convolutional 2D layer coupled with the tangent activation function wrapped into a spiking layer.

The discriminator model comprises two stacks of convolutional 2D layers followed spiking LeakyReLU activation layers with the slope of 0.2 and a final convolutional 2D layer.

The hyperparameters included Adam optimiser with the learning rate of 0.0001 and BCEWithLogitsLoss as a loss function for both models. Batch size was equal to 64 data examples, the time steps (an important hyperparameter for spiking layers) for a forward pass was set to a value of 100 in both – the generator and discriminator models, the number of training epochs totalled to 200. The weights of the convolutional layers were initialised from a zero-centered normal distribution with standard deviation 0.02.

### 4.3 Training results and discussion

The very first prototype of the network doesn’t converge yet well enough, with the generator being inferior in performance comparing to the discriminator. There are therefore a room for amelioration which might be accomplished with the application of the surrogate backpropagation methods (spiking aware backpropagation) and a plethora of conventional GAN stabilization methods, such as earth mover’s distance algorithm, Lipschitz continuity and more recent regularization methods (Lee & Seok 2020), since the semantic music modality being transformed to a visual modality with the proposed encoding technique, may benefit from the discoveries made for GANs in a visual domain.

### 5. SUMMARY

The article presented an adaptive VR test in music harmony, based on the original representation methodology, employing colour and colour shades for harmonic function and chord structure representation respectively, lifting the visual cognitive load for the learner while auditory adoption of new chords and harmonic sequences. The generative mechanism of the test is based on conditional convolutional spiking GAN for semantic music generation. Novel encoding strategy allowed transforming one-dimension array of the sequence into two-dimensional representation, using the position of each chord in a system of graphs in music harmony, which allowed a considerable data augmentation. The work presents a significant step in research of spiking neural network paradigms application to the problematics of semantic music generation.
6. REFERENCES


---

**Adaptive VR Test in Music Harmony Based on Conditional Spiking GAN**

Anna Shvets & Samer Darkazanli


Adaptive VR Test in Music Harmony Based on Conditional Spiking GAN
Anna Shvets & Samer Darkazanli


