REDES NEURAIS E SÍNTESE MUSICAL UTILIZANDO CONJUNTO DE DADOS SONOROS

NEURAL NETWORKS AND MUSIC SYNTHESIS USING SOUND DATA SETS

Gabriel Francisco Lemos
Research Center on Sonology, University of São Paulo, USP
Group on Artificial Intelligence and Art, GAIA, InovaUSP
São Paulo, Brazil
gabriel.lemos@usp.br

Resumo. O presente artigo propõe um estudo comparativo entre duas estruturas topológicas de Redes Neurais – *Recursive Neural Networks* (RNN) e *WaveNet* – aplicadas à síntese sonora e análise de conjunto de dados sonoros. Avaliou-se o estado da arte dessas tecnologias no campo da criação sonora contemporânea, identificando seus limites técnicos e possibilidades estéticas na aplicação desses sistemas em contextos artísticos. A relevância da pesquisa na implementação desses modelos no campo da criação sonora e no contexto brasileiro se concentra no estudo crítico da adequação das técnicas de aprendizado de máquina na síntese e nas implicações estéticas para a composição contemporânea. No atual estágio da pesquisa, concluímos que a aplicação desses métodos de síntese se encontra aquém de uma utilização profissional, visto que os sons produzidos possuem alto índice de ruído, apresentam baixa resolução e dificilmente mantêm uma coerência composicional no decorrer do tempo das amostras. Ressaltamos também que a implementação desses sistemas no contexto brasileiro é problemática, pois o desenvolvimento desses modelos necessita de acesso a custosos recursos computacionais de alto desempenho. Identificamos, no entanto, que uma alternativa possível para esse problema de acesso às infraestruturas adequadas é a contratação de serviços de processamento via nuvem – mas que, salientamos, são monopolizados por companhias localizadas exclusivamente no Norte Global.

Palavras-chave: Redes neurais; Aprendizado de Máquinas; Processos Criativos; Curadoria de conjunto de dados; Síntese Sonora.

Abstract. This article proposes a comparative study between two topological structures of Neural Networks – Recursive Neural Networks (RNN) and WaveNet – applied to sound synthesis and analysis of sound datasets. Based on these two specific systems, the state of the art of these technologies in the field of contemporary sound creation was evaluated so that it could be possible to identify technical limitations and aesthetic possibilities for applying these systems in musical contexts. The relevance of the research in implementing these models in the field of sound creation and the Brazilian context focuses on the critical study of the adequacy of machine learning techniques in synthesis and the aesthetic implications of this technology in composition practices. At the current research stage, we conclude that the application of these synthesis methods falls short of professional use since the sounds produced have a high noise index, have low resolution and hardly maintain compositional coherence over the time of the samples. We also emphasize that implementing these systems in the Brazilian context is problematic, as developing these models requires access to costly high-performance computational resources. Hence, we have identified that a possible alternative to this problem of access to adequate infrastructure is the subscription to processing services via the cloud – we emphasize, however, that they are part of a monopoly of technology companies located exclusively in the Global North.

Keywords: Neural networks; Machine Learning; Creative Processes; Dataset curation; Sound Synthesis.

INTRODUÇÃO

Desde a primeira década do Século XXI, em diversas disciplinas, setores comerciais e áreas acadêmicas é notável um crescente interesse nas chamadas tecnologias de Inteligência Artificial (IA). Nota-se, mais especificamente, um grande investimento e avanço em pesquisas de aprendizado de máquina [*machine learning* - ML] dedicadas ao processamento de grandes quantidades de informação digital. Na última década, um fator que influenciou diretamente esse movimento foi a ampliação, o desenvolvimento e a disponibilização significativa de softwares e bibliotecas para linguagens de programação específicas – dois exemplos paradigmáticos são as plataformas *TensorFlow 2* (GOOGLE a, 2021) e o *PyTorch* (FACEBOOK, 2021). Em larga escala, tal disponibilidade de meios computacionais tornou o desenvolvimento no campo do aprendizado de máquinas mais acessível para usuários não especializados. A onipresença da linguagem de programação *Python* também não pode ser ignorada e deve ser mencionada como um fator que contribuiu nesse rápido desenvolvimento. Se vista em comparação com linguagens como o *C* + +, a popularização do

Python se dá tanto devido à sua acessibilidade, quanto à sua relativa facilidade de implementação em sistemas distintos.

Atualmente, a retomada do otimismo[1] em pesquisas e aplicações de Inteligência Artificial se recompôs a partir da retomada de arquiteturas computacionais chamadas de redes neurais, assim como de um rápido avanço do paradigma *deep learning* – abordagem de aprendizado de máquinas que considera a IA como a habilidade de agentes computacionais em "aprender" pela experiência[2] ao analisar e processar um grande conjunto de dados digitais. A retomada de interesse na área pode ser atribuída ao aumento da capacidade de generalização dos modelos atuais, o crescimento da nuvem como serviço de armazenamento e computação de dados e a acessibilidade comercial à componentes de hardware avançados – como às GPUs (*Graphic Processing Units*). O otimismo da indústria também cresceu significativamente devido ao surgimento de soluções de software que tornaram as arquiteturas de Redes Neurais mais acessíveis e fáceis de programar (vide os serviços de processamento oferecidos através da plataforma *Google Colab*[3]).

O recente "boom" do aprendizado de máquinas também levou a um interesse expandido no chamado *creative programming* – processos computacionais que focam na criação de algo expressivo e inovador ao contrário de um uso exclusivamente utilitário. Em muitas práticas artísticas contemporâneas, em especial de produção e composição musical (foco do presente artigo), a utilização de sistemas computacionais se tornou indissociada dos processos criativos em multimídia (SCHUBERT, 2021). Na produção artística independente ou comercial, com o advento do aprendizado de máquinas aplicado à síntese sonora, a relação entre criação de conteúdo musical e manipulação do áudio digital se tornou particularmente fecunda devido a habilidade desses sistemas automatizados gerenciarem uma grande quantidade de amostras sonoras organizadas em bancos de dados personalizados e curados para fins específicos – aumentando exponencialmente a capacidade de sistemas generativos em sintetizar longas horas de áudio, após um período de modelagem de uma determinada rede neural.

Nesse contexto, o presente artigo visa introduzir alguns dos conceitos e processos técnicos fundamentais para esse tipo de processamento sonoro, também conhecido como *neural audio synthesis*. Focaremos, especialmente, na relação entre os processos de síntese via redes neurais e alguns dos conhecimentos compartilhados com a ciência da informação, como por exemplo a curadoria de conjunto de dados sonoros, a visualização multidimensional da informação musical e a modelagem estatística realizada em ambientes computacionais programados via comando de texto.

METODOLOGIA

A pesquisa se caracteriza como bibliográfica e documental baseada em metodologia experimental e analítica organizada em torno de experimentos generativos aplicados, marcos e testes sucessivos na implementação de redes neurais em contexto artístico sonoro. A implementação computacional da pesquisa se vale de procedimentos metodológicos qualitativos e quantitativos de tipo exploratório experimental em ambiente *Python* – utilizando bibliotecas da plataforma *TensorFlow 2* e executados no *Google Colab* utilizando uma GPU Tesla T4.

Tendo em vista uma introdução ao campo interdisciplinar da síntese sonora por meio de aprendizado de máquinas, o presente artigo propõe uma revisão bibliográfica na área do desenvolvimento das primeiras redes neurais aplicadas à síntese e à visualização multidimensional de dados complexos (ZHANG, 2008). Com o intuito de ilustrar os processos computacionais analisados no decorrer do texto, quando necessário, apresentamos algumas representações visuais e exemplos sonoros realizados pelo próprio autor e terceiros. Ao abordar essa área de cruzamento entre as práticas artísticas e o gerenciamento de dados, priorizamos comentar os desenvolvimentos tecnológicos da última década, observando as estruturas topológicas de duas

[1] No entanto, em retrospectiva, o nascimento da *World Wide Web* também coincidiu com o início do chamado "Inverno da Inteligência Artificial" (NORVIG e RUSSELL, 2021, p.74), diversas pesquisas que levavam a sigla da "IA" seriam interrompidas após uma década de exageros, seguida de eventuais fracassos em corresponder às expectativas de seus investidores.

[2] Os modelos ditos conexionistas eram vistos por alguns cientistas como concorrentes diretos tanto dos modelos simbólicos promovidos por Allen Newell (1927-1996) e Herbert Simon (1916-2001), quanto pela abordagem lógica de John McCarthy (1927-2011). Segundo Stuart Russell e Peter Norvig em *A Modern Approach to Artificial Intelligence*, "Pode ser que os modelos conexionistas formem conceitos internos de uma forma mais fluida e imprecisa, portanto mais adequada à confusão do mundo real. Eles também têm a capacidade de aprender com os exemplos – eles podem comparar seu valor de saída previsto com o valor real em um problema e modificar seus parâmetros para diminuir a diferença, tornando-os mais propensos a ter um bom desempenho em exemplos futuros." (NORVIG e RUSSELL, 2021, p.75).

[3] O *Google Colaboratory* é um serviço que possibilita o acesso ao ambiente de programação (*notebooks*) *Jupyter* que não requer configuração prévia que utiliza processamento de nuvem. Ele permite que o usuário escreva e execute códigos em *Python*.

redes neurais consolidadas, os sistemas *WaveNet* e *SampleRNN*. No atual estágio da pesquisa, visamos identificar como as mudanças estruturais na organização de camadas dessas redes se relacionam e influenciam diferentes modos de manipulação e reconhecimento de padrões na informação do áudio digital analisado.

DESENVOLVIMENTO

Por vezes, o termo Inteligência Artificial é citado na imprensa, bem como por artistas e formadores de opinião, como uma espécie de termo "guarda-chuva" ou processo mágico (VICKERS e ALLADO-MCDOWELL, 2021; STEYERL, 2017, p.47) cujo funcionamento técnico é raramente apresentado para o público não especializado. Almejando um efeito contrário, buscou-se nesse artigo esclarecer a relação entre o desenvolvimento de IA e as pesquisas em torno do aprendizado de máquinas com Redes Neurais Artificiais em contextos sonoro musicais.

Redes Neurais Artificiais

Uma rede neural (*Neural Network*, ou NN) é um sistema feito de "neurônios artificiais", unidades computacionais com entrada e saída que quando modeladas em uma arquitetura iterativa complexa, busca simular o comportamento de agentes autônomos que demonstram aptidões de aprendizado (NORVIG e RUSSELL, 2021, p.1378). As primeiras estruturas desse tipo datam do final da década de 1950 e foram desenvolvidas em sistemas especializados por laboratórios de pesquisa em universidades do Norte Global. As redes neurais são construídas a partir de estruturas topológicas compostas por várias camadas, cada uma contendo um ou mais "neurônios". Um nível dessa topologia (ou *tier,* conforme usado na implementação desses sistemas) pode encapsular várias camadas internas ou "ocultas", cada uma consistindo em várias centenas, ou talvez até milhares de "células". Esses "neurônios artificiais" são estruturas tipicamente muito simples, sistemas de entrada-saída aos quais são adicionados pesos numéricos arbitrários, ou aleatórios, que refletem vieses[4] algorítmicos.

O *perceptron*, a forma discreta mais básica presente numa rede neural clássica, consiste em uma unidade ou "neurônio". Como formalizado matematicamente pela Figura 1, um *perceptron* apresenta quatro entradas separadas (**a**n), cada uma multiplicada por um peso (**W**n) que reflete uma relação direta com as necessidades de determinado modelo computacional. Normalmente, um valor de unidade de polarização é adicionado à soma destes e o resultado é multiplicado na saída por uma função de ativação (**σ**) cujo valor determina se o neurônio "dispara" ou não – a ativação dessas unidades distribuídas em diversas camadas da rede é o que resulta no modelo da rede neural utilizada.

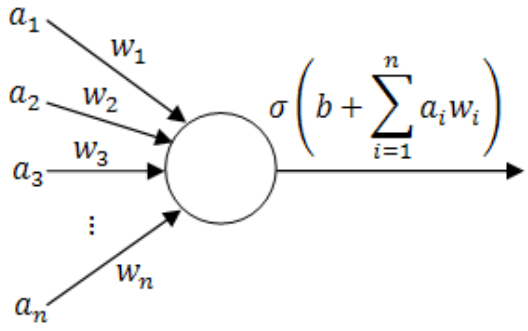

Figura 1. Representação matemática de um *Perceptron.*
Fonte: PERCEPTRON (2011).

Conjunto de Dados Sonoro-Musicais como entrada

Para que uma rede neural aprenda, ela deve ser treinada, o que envolve expor repetidamente a estrutura de camadas do sistema a um conjunto de dados [*dataset*] de treinamento. Tal processo é trabalhoso e pode

[4] Ressaltamos que o termo viés, comumente vinculado à uma conotação negativa quando analisado em contextos sociais mais amplos, não é necessariamente um problema de funcionamento do ponto de vista dos desenvolvedores dessas tecnologias. Na perspectiva da ciência da computação e da estatística, o viés algorítmico é um *feature*, não um *bug* desses sistemas. No entanto, de um ponto de vista ético mais inclusivo, ressaltamos que o viés é um sintoma dos vetores epistêmicos e das relações de poder atuantes no campo da computação e da indústria tecnológica atual.

levar várias horas, ou mesmo dias, dependendo do tamanho e da complexidade dos dados de entrada. Nessa etapa de elaboração dos bancos de dados é comum analisar as características dos elementos que compõem tal *dataset*. Diferentes métodos de análise e visualização podem ser utilizados para melhor compreender a complexidade da informação que compõem esses conjuntos de dados. No Gráfico 1, composto pelo algoritmo t-SNE (*t-Distributed Stochastic Neighbour Embedding)*[5], aplicamos um método de visualização que descreve arquivos de áudio em dois *datasets*: uma gravação do campo eletromagnético (DVS SOUND, 2017) produzido por um carro elétrico (imagem superior) e a composição *Binah* (LEMOS, 2016) para percussão múltipla (imagem inferior). A visualização pelo t-SNE propicia uma redução de dimensionalidade com o objetivo de representar a alta variação de atributos dos arquivos sonoros (MAATEN e HINTON, 2008). Leon Fedden (FEDDEN, 2017) aborda com mais detalhes este método aplicado ao reconhecimento de padrões em arquivos de áudio digital.

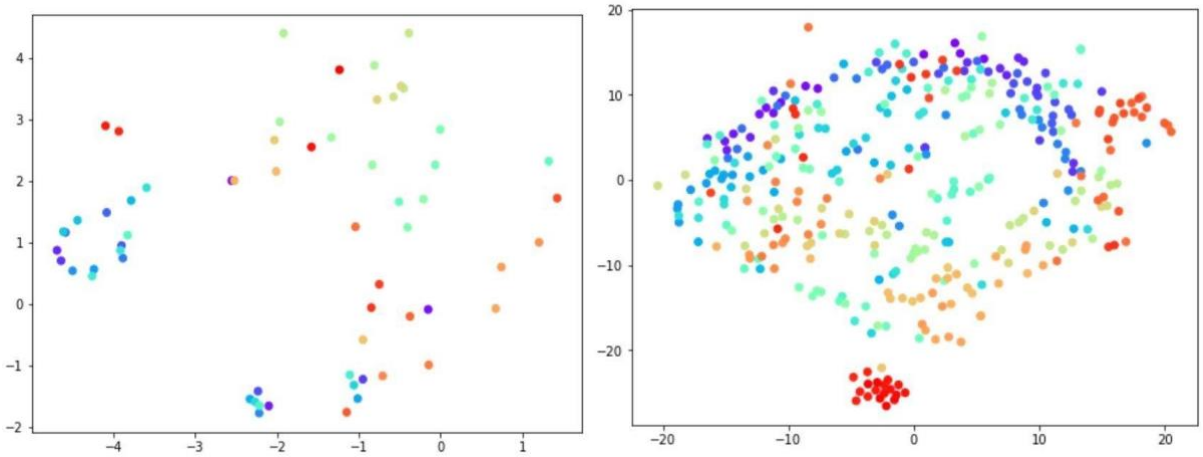

Gráfico 1. Visualização de áudio via t-SNE. O eixo vertical indica atributos frequenciais na banda espectral analisada e o eixo horizontal representa variações de amplitude no tempo da gravação original.
Fonte: Imagem produzida pelo autor (2021).

Como indicado no espaço latente[6] do Gráfico 1, sons parecidos tendem a estar a uma distância menor entre si, ao mesmo tempo que também são representados pelas mesmas cores. Pode-se notar que o Gráfico 1 (imagem superior) tem uma menor diversidade de elementos (caracterizada por sua esparsidade). O Gráfico 1(imagem inferior) tem uma maior variedade espectral e densidade de atributos.

Ressaltamos também uma clara polarização em determinadas regiões do espectro – o grave referente ao Bumbo e aos Tímpanos (em vermelho) sendo um exemplo indicativo desse padrão. Com o auxílio desse tipo de visualização, podemos deduzir que o processo de aprendizagem de uma dada rede neural será mais custoso e demorado se posta a analisar o conjunto mais denso.

Treinando redes neurais

Ao longo de várias "épocas" [*epochs*] ou passagens completas pelo conjunto de dados curado pelo usuário, os pesos e vieses algorítmicos de cada camada que compõem a rede em questão são ajustados, seja automaticamente ou "afinados" de acordo com as necessidades do programador. Historicamente, o processo de aprendizado de máquinas via redes neurais tem um precursor no conceito matemático de "regressão linear" (NORVIG e RUSSELL, 2021, p.676). Representada no Gráfico 2, a regressão linear envolve encontrar um "vetor de ajuste ótimo" para um conjunto de dados definido. Durante o processo de computação transcorrido nessa etapa, o nível de "aptidão" [*accuracy value*] é medido usando uma função de perda [*loss value*] que, por sua vez, retorna o erro entre a etapa de treinamento mais recente e o coeficiente inicial. Idealmente, segundo Norvig e Russel, a função de perda deve ser menor no final do treinamento do que no início, esse processo de otimização também é conhecido como "redução de perda" (idem, p.1386).

[5] O t-SNE oferece uma representação bidimensional a partir de uma PCA (*Principal Component Analysis*) das amostras de áudio. O eixo vertical indica atributos no campo das frequências na banda espectral e o eixo horizontal representa variações de amplitude (volume). O zero em cada eixo corresponde ao ponto de convergência (centróide) dos três planos de análise do PCA.
[6] Um espaço latente é um espaço multidimensional abstrato que contém valores de características que não podemos interpretar diretamente, mas que pode codificar uma representação interna significativa de eventos observados externamente. O que poderia ser encarado como um espaço de incorporação, em que um conjunto de itens dentro de um coletor são representados em relação a itens semelhantes entre si e posicionados mais próximos uns dos outros – formando assim o espaço latente.

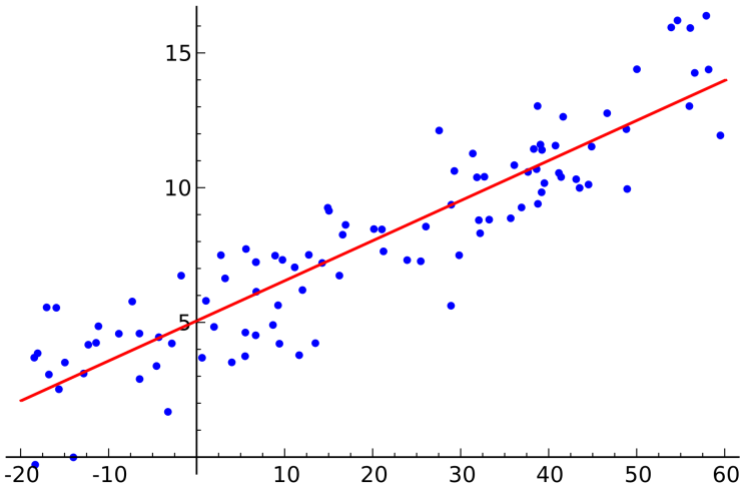

Gráfico 2: Representação gráfica de uma regressão linear. Regressão linear, mostrando um conjunto de dados aleatórios em que a linha vermelha corresponde ao ajuste ótimo da informação analisada.
Fonte: Wikipedia (2021).

Uma vez que a rede tenha sido treinada, o "modelo" resultante terá, idealmente, identificado a recorrência dos padrões que compõem os elementos do conjunto de dados inicial. Esse modelo refletirá a pregnância informacional dos dados analisados e poderá ser salvo, classificado e aplicado posteriormente para sintetizar digitalmente novas amostragens, ou para modificar arquivos de áudio existentes[7]. Em suma, esse tipo de síntese ou transformação pretende mimetizar estatisticamente a recorrência de padrões encontrados no conjunto de dados analisado durante a etapa de treinamento. Nesse sentido, o modelo produzido por meio de redes neurais pode ser usado no sampleamento digital de diversos arquivos de áudio (.wav) que, por sua vez, possibilitam replicar globalmente a multidimensionalidade sonora do *dataset* – idealmente, buscando semelhanças de intensidade, timbre, padrões rítmicos, melódicos, etc.

Redes Neurais aplicadas à Síntese de Novos Sons

Dado o sucesso mostrado pelas redes neurais na predição de sequências e geração de texto (HOCHREITER e SCHMIDHUBER, 1997; GRAVES, 2013; KARPATHY, 2015), constatou-se que os mesmos processos podem ser aplicados para a síntese de outros tipos de informação sequencial, como por exemplo, geração de regras composicionais e o sampleamento digital no domínio da forma de onda sonora. No entanto, o aprendizado de máquinas via redes neurais só ganhou ampla notoriedade no contexto do áudio a partir de 2015[8] com a publicização de técnicas e modelos específicos como o *WaveNet* (VAN DEN OORD et al., 2016), *SampleRNN* (MEHRI et col, 2016), *Nsynth* (ENGEL e RESNICK, et al., 2017), *DeepVoice* (ARIK et al. 2017), *WaveRNN* (KALCHBREMER, et al., 2018), *GANSynth* (ENGEL et al., 2019), o *Music Transformer* (HUANG, et al., 2018), entre outros. Na Figura 2, feita por Fjodor Van Veen (VEEN, 2016), podemos observar uma simplificação gráfica que esquematiza as diferentes camadas que compõem as topologias de redes neurais presentes nos modelos mais comuns na área: as *Variational Auto-Encoders* (VAE) das *Generative Adversarial Networks* (GAN) e as *Long Short Term Memory* (LSTM) e *Gated Recurrent Unity* (GRU) atuantes na *Recursive Neural Networks* (RNN).

---

[7] Especialmente presente em processos artísticos ligados à visualidade e à experimentação sonora, esse tipo de prática pode ser aplicado por meio da técnica de *Style Transfer*. Em 2014, no campo da produção visual, essa estética ficou conhecida como *Deep Dreams* ou *Inceptualism* (GOOGLE b, 2021). Hoje em dia resultados semelhantes podem ser alcançados utilizando o sistema *StyleGAN2-ADA* (SCHULTZ, 2021). No campo da música, esse processo de transformação é possível por meio de diferentes técnicas. Dentre elas, ressaltamos o modelo produzido pela cantora Holly Herndon, que pode ser experimentado via *browser* (HERNDON, 2021) e o R.A.V.E. (*Realtime Audio Variational autoEncoder*) (CAILLON e ESLING, 2022).

[8] Aqui, vale a ressalva que a relação entre IA e música não é recente. Pontuo apenas que entre 2015 até hoje a acessibilidade a essas tecnologias ganhou maior abrangência. Desde os anos 1950 pesquisas são feitas na tradução simbólica entre modelos probabilísticos e composição musical. O projeto *Illiac Suite* (mais tarde renomeada como *String Quartet No. 4*) é paradigmático nesse aspecto, composta em 1957 para quarteto de cordas, a obra é considerada a primeira partitura escrita por um computador eletrônico. Lejaren Hiller, em colaboração com Leonard Issacson, programou o computador ILLIAC I na Universidade de Illinois em Urbana – Champaign (onde ambos os compositores eram professores).

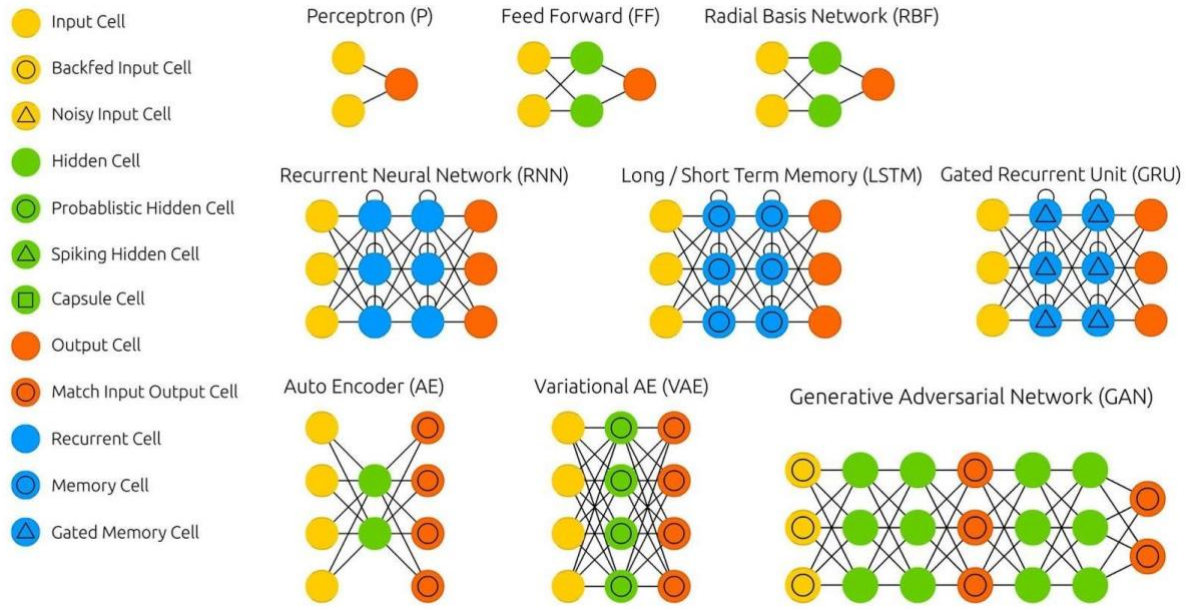

Figura 2. Representação gráfica das topologias entre tipos de Perceptrons, RNN e GAN.
Fonte: VEEN (2016).

As representações topológicas feitas a partir de camadas e nós apresentam uma ferramenta de visualização comparativa entre as arquiteturas mais conhecidas. Por meio de variações gráficas econômicas é possível indicar que a RNN, por exemplo, pode ser implementada com alterações no processamento sequencial, sejam elas pela implementação de camadas "mnemônicas" de tipo *Long Short Term Memory* (LSTM) ou *Gated Recurrent Unit* (GRU). As GANs, por sua vez, incorporam em sua arquitetura processos de compressão que produzem um espaço latente de variações possíveis organizadas via *Auto Encoder* ou *Variational Auto-Encoder*.

Nessa área de pesquisa é importante salientar a existência de dois campos de desenvolvimento em música: um aplicado ao processamento de sinais de áudio no domínio da forma de onda (o som como representação digital do fenômeno acústico propriamente dito) e o outro, relacionado a representação simbólica de tais sinais – como texto escrito, protocolo MIDI ou partitura musical. Várias tentativas foram feitas para modelar este último, incluindo diversas aplicações comerciais como o *Aiva* (2020) e o projeto *Magenta Studio* (ECK, 2016) da *Google* – todos focados no processamento e geração de arquivos em formato MIDI. Nossa pesquisa, no entanto, se concentra no primeiro tipo de modelo gerativo em que, rescindido os dados simbólicos, utilizamos sinais de áudio digital no domínio da representação da forma de onda. Ou seja, testou-se sistemas de manipulação informacional que recebem sinais de áudio digital e consequentemente, geram novos sinais de áudio digital em nível acústico (pronto para serem ouvidos). No recente desenvolvimento desse tipo de processamento, existem alguns sistemas *end-to-end*, os quais os dois mais amplamente divulgados são os já mencionados *WaveNet* e o *SampleRNN*.

Modelo Wavenet

O *WaveNet* é um sistema desenvolvido pelo projeto *DeepMind* (com investimento do *Google* desde 2014). Originalmente desenvolvida no contexto da síntese de texto para voz (ou *text-to-speech synthesis,* TSS), a *WaveNet* foi adotada em contextos artísticos como uma ferramenta desenvolvida para o mercado formado por músicos e compositores. Ela utiliza uma categoria específica de rede, a chamada Rede Neural Convolucional (CNN ou *ConvNet*). Muito conhecida no campo do processamento de imagens e visão computacional, essa aplicação obteve resultados pioneiros, porém, eticamente problemáticos ao realizar tarefas de classificação de imagens e detecção de elementos visuais com altas ocorrências de racismo e desigualdade/discriminação de gênero (AMOORE, 2020; BROUSSARD, 2018; EUBANKS, 2018; GRAY e SURI, 2019; MALIK, 2020).

Como mostra a Figura 3, a operação básica de uma CNN envolve a aplicação de um ou mais filtros na superfície bidimensional de uma imagem de entrada. Por meio de um processo de multiplicação e somatória,

esses filtros (ou *kernels*) são capazes de redimensionar os vetores de *pixel* em diferentes partes da imagem. Para a geração de áudio, por ser fundamentalmente pautada em sequências temporais, a dimensionalidade é reduzida de 2 para uma única dimensão temporal – no entanto, o princípio básico e a operação computacional permanecem os mesmos.

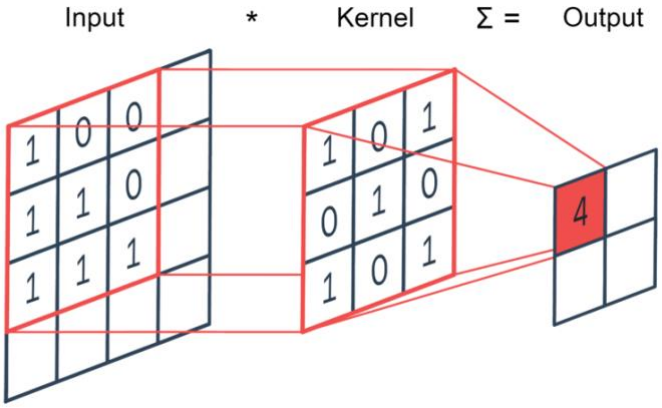

Figura 3: Convolução de redução informacional 2D. Cada célula na matriz de entrada é multiplicada pela célula correspondente no *kernel*. Os novos valores são somados, o resultado representando uma única célula de saída, gera uma matriz de duas dimensões. Todo o processo é repetido com o *kernel* deslocado para a direita em uma ou mais etapas até que toda a matriz de entrada tenha sido processada.
Fonte: MELEN (2020).

O modelo gerador da *WaveNet* é composto por várias camadas de convolução, estes se combinam para produzir uma distribuição probabilística da próxima amostragem – considerando como parâmetro condicional todas as amostras analisadas anteriormente. A *WaveNet* usa uma forma modificada de convolução que envolve um processo chamado dilatação (ou "convolução com buracos") em que, as saídas das sucessivas camadas ocultas da rede são "podadas", causando com que algumas saídas sejam omitidas quando também forem aplicadas como possíveis entradas na próxima camada da rede. Como evidente na Figura 4, a cada nova camada, a entrada anterior é dilatada. O aumento exponencial dessas dilatações permite com que o sistema se expanda rapidamente gerando milhares de amostras em sequência temporal.

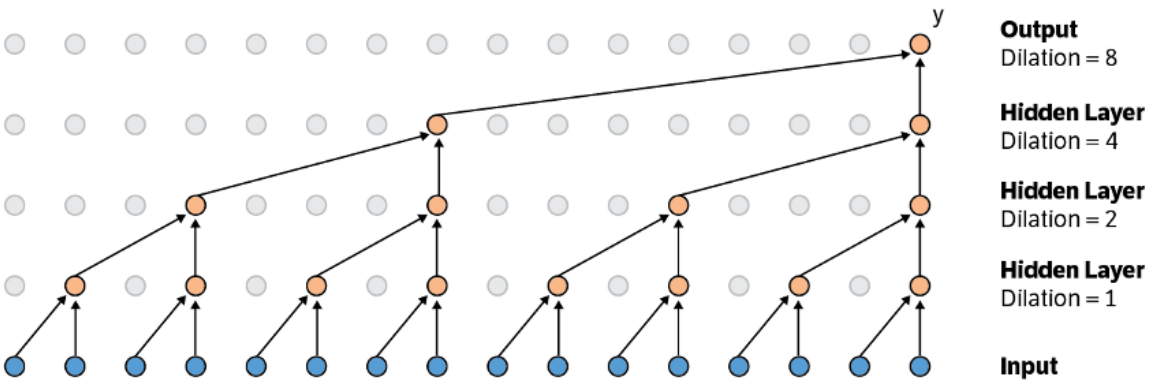

Figura 4: Processo de dilatação *Wavenet*.
Fonte: VAN DEN OORD et al. (2016).

A *WaveNet* tem se mostrado bem-sucedida na produção de resultados sonoros de relativa alta qualidade, especialmente na área da síntese de voz. Ela sem dúvida teve um menor sucesso ao modelar estruturas musicais de durações superiores a alguns segundos – o que não quer dizer que seja incapaz de produzir sequências musicais estendidas, mas sim, que sua coerência estrutural tende a diminuir à medida que as escalas de tempo aumentam (HINER, 2019). Apesar de existirem concorrentes competitivos como a *Jukebox* da *OpenAI* (DHARIWAL, 2020), a *WaveNet* continua sendo o modelo referencial para síntese de áudio digital usando redes neurais. No entanto, ressaltamos que a natureza condicional dessa arquitetura –

em que cada amostra gerada depende daquelas das etapas anteriores – acarreta em um grande impacto de desempenho durante o treinamento e geração de áudio[9]. Como a rede é essencialmente pautada em distribuições binárias sequenciais, seu custo computacional global é muito alto, tornando o sistema impraticável em muitos computadores que não possuem GPUs potentes.

Modelo SampleRNN

Semelhante a *WaveNet*, a *SampleRNN*[10] processa diretamente as amostras de áudio digital (.wav), mas difere na arquitetura utilizada. Enquanto a *WaveNet* é baseada em redes convolucionais, a *SampleRNN* utiliza um tipo conhecido como Rede Neural Recorrente, ou *Recurrent Neural Networks* (RNN). Essas redes foram desenvolvidas para processar dados sequenciais, como séries temporais ou qualquer informação que possa ser modelada de forma sequencial, como texto e informações de áudio – a fala ou a sintaxe musical de uma melodia, por exemplo. Ao contrário de redes simples de tipo *Feed-Forward* (como o *perceptron*), as RNNs retêm uma espécie de memória interna que registra seus estados anteriores (daí o rótulo "recorrente"). Tal qual esquematizado na Figura 5, uma RNN inicia seu funcionamento quando o estado de saída de uma unidade é alimentado como informação de entrada na etapa seguinte.

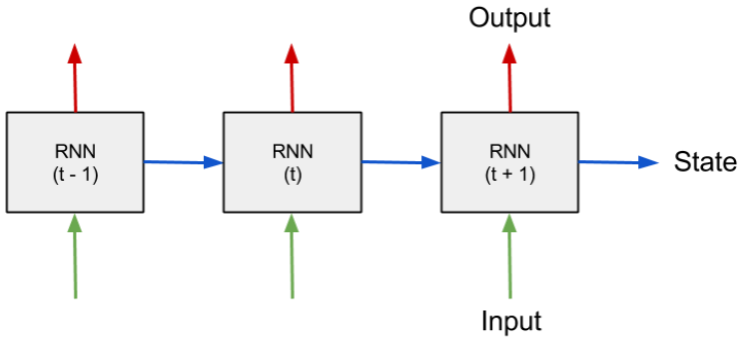

Figura 5. Rede Neural Recorrente.
Fonte: MELEN (2020).

Observe que esta estrutura é uma simplificação, uma vez que a RNN consiste em uma única unidade que itera diversas vezes ao longo de muitas épocas de treinamento – como em um *loop* recorrente que, por sua vez, alimenta de volta sua própria saída a cada novo estado do sistema. Além de "memorizarem" por meio dos *Long Short Term Memory* (LSTM) ou *Gated Recurrent Unit* (GRU), essas redes também são capazes de "esquecer" dados aprendidos. Nesse caso, estados posteriores a um certo limite de tempo configurável são passíveis de serem apagados automaticamente do histórico de aprendizado da rede. Como outras redes neurais, a saída do modelo *SampleRNN* representa uma distribuição probabilística em que a próxima amostra gerada é fruto de um processo preditivo que identifica a "pregnância informacional" emergente nos padrões aprendidos na análise do conjunto de dados introduzidos no sistema.

Resultados

Como observamos brevemente ao discutir a *WaveNet*, um dos principais problemas na modelagem de sinais de áudio digital é a questão da escala ou dimensionalidade dos dados analisados e futuramente sintetizados. Comumente, as estruturas que constituem um sinal de áudio digital operam simultaneamente a partir de diferentes parâmetros (altura, duração, intensidade, timbre, aspectos composicionais, etc.) e representações digitais desses parâmetros (espectrogramas, *Self-Organizing Maps*, PCAs, etc.) são recorrentes. Como indicado na representação visual do algoritmo t-SNE (Gráfico 1), amostras de áudio possuem relações acústicas com outras amostras vizinhas ou até mesmo separadas por vários períodos de tempo. Um

[9] Uma implementação feita a partir da biblioteca *TensorFlow* pode ser encontrada em: <https://github.com/ibab/tensorflow-wavenet>. Acesso feito em 25 de agosto de 2021.
[10] A arquitetura *SampleRNN* foi publicamente apresentada no artigo *SampleRNN: An Unconditional End-to-End Neural Audio Generation Model* (MEHRI *et al.,* 2017). Os autores lançaram uma implementação disponibilizada no GitHub. No entanto, atualmente o código não está em desenvolvimento ativo e depende de vários pacotes de software desatualizados (incluindo o Python 2.7, sem suporte desde início de 2020), o que o torna obsoleto. Uma versão atualizada pode ser encontrada no modelo *Prism-SampleRNN* (SALEM, 2021)

problema que surge a partir desse tipo de relação é a discrepância entre a dimensionalidade do sinal de entrada de áudio no domínio da forma de onda e sua amostragem sequencial na saída do sistema computacional utilizado. Mesmo com uma taxa de amostragem baixa, como 11025 Hz, precisaremos produzir milhares de amostras antes que qualquer som reconhecível ou esteticamente interessante seja produzido[11]. Modelar dependências de longo prazo a partir do *dataset* de entrada é, portanto, inerentemente problemático do ponto de vista técnico e, por sua vez, introduzem limites técnicos e estéticos no uso desse tipo de síntese em contextos artísticos que exigem uma qualidade de áudio de nível profissional.

Para mitigar esse tipo de problema, o *SampleRNN* adota uma arquitetura organizada hierarquicamente que consiste em redes recorrentes separadas por camadas construídas "umas sobre as outras" e com pesos distintos – ou seja, cada camada apresenta uma resolução mais ampla do que a camada imediatamente abaixo dela. Quadros de amostragem de tamanho fixo são consumidos em cada camada, a mais baixa sendo resolvida no nível de amostras individuais – a camada mais baixa não é de fato uma RNN, mas sim um *Perceptron Multi-Layer* (MLP), uma espécie de rede *feedforward* mais simples. Cada camada, exceto a mais alta, também é condicionada pela saída da camada imediatamente acima. Tendo esse tipo de configuração hierárquica, as redes de tipo *SampleRNN* podem ser usadas para gerar sons de duração indeterminada – ou circunscrita ao limite proporcionado pela memória do computador utilizado.

Embora seja possível executar os *scripts* de treinamento e geração de áudio apenas utilizando uma única GPU treinada a partir de um conjunto de dados – digamos um álbum de uma hora de duração – um desempenho ideal (ou mesmo aceitável) exigirá métodos para gerar dados aumentados (*dataset augmentation*). No contexto do aprendizado de máquinas, o aumento no conjunto de dados é uma técnica que pode ser usada para expandir artificialmente o tamanho de um *dataset*, criando versões modificadas e variações nas amostras iniciais – o que melhora o desempenho e a capacidade de generalização do modelo. Técnicas usadas em dados sonoros podem significar a inclusão de ruídos mixados ao áudio original, mudança na velocidade de reprodução das amostras (*speed shifting*), a alteração no ponto inicial de reprodução desses arquivos (*time stretching*) e a mudança da frequência (nota) das amostras (*pitch shifting*). Observamos que épocas de treinamento feitas em *datasets* não aumentados apresentavam insuficiência no aprendizado, impossibilitando uma modelagem mais precisa e uma síntese aceitável.

Infelizmente, até o momento da publicação deste artigo, a qualidade de áudio produzida pelos experimentos com a *SampleRNN* se manteve abaixo do padrão profissional. Apesar da potencialidade das redes neurais apresentadas aprenderem a partir de um grande conjunto de dados sonoros e de produzirem incontáveis horas de áudio digital, foi observado que a síntese musical com esses sistemas ainda apresenta uma sonoridade com fortes traços de cancelamento de fase, som de resampleamento, altos índices de ruído indesejado e baixa coerência composicional em longos períodos de tempo – especialmente no delineamento melódico das amostras. No entanto, salientamos que um relativo grau de inventividade e imprevisibilidade rítmica pode ser alcançado.

Sobre esse potencial inventivo, o modelo *SampleRNN* ganhou notoriedade em 2019 por meio da publicação do *Relentless Doppelganger* (DADABOTS, 2019), criada pela dupla Dadabots[12]. A pesquisa musical da banda – formada pelos músicos e pesquisadores CJ Carr e Zack Zukowski – baseia-se em processos de síntese que necessitam de alto desempenho computacional para proporcionar um fluxo contínuo de som sintetizado a partir de gêneros musicais específicos. Desde então, uma série de experimentos gerativos de tipo *neural synth* (CARR e ZUKOWSKI, 2017; 2018; 2019) foram feitos pela dupla utilizando o sistema *SampleRNN* – vide *Neural Black Metal* (2017), *Neural Death Metal* (2019) e *Neural Free Jazz* (2019), dentre outros (DADABOTS, 2021).

CONSIDERAÇÕES FINAIS

Neste artigo, examinamos uma das áreas mais publicizadas da atual pesquisa na intersecção entre música, redes neurais e a curadoria de dados sonoros. Comparamos dois dos modelos mais consolidados nessa área – a *WaveNet* e a *SampleRNN*. Além das capacidades oferecidas no processamento de ambas, certo otimismo na implementação de aprendizado de máquinas em áudio digital se apresenta na divulgação das *Generative*

[11] Para uma demonstração dessa etapa de treinamento e resultados preliminares na pesquisa em síntese, ouça os exemplos das 165 épocas de treinamento de uma rede neural de tipo *SampleRNN* treinada a partir de um conjunto de dados composto por duas horas de gravações de peças para piano compostas por Chiquinha Gonzaga (1847-1935) e interpretadas pela pianista Maria Teresa Madeira (1960-): <https://drive.google.com/drive/folders/1xSrIAyUNQvielCvBW7n0rjYTqqBrR4Zy?usp=sharing>.

[12] Dadabots lançou seu próprio código no GitHub, mas como ele é baseado na base de código Python 2 original em vez de ser uma nova implementação, atualmente é um desafio colocá-lo em funcionamento.

Adversarial Networks (GANs) – uma arquitetura de rede neural que também tem se mostrado um sucesso no campo da síntese de imagens e vídeos digitais.

Com a criação de sistemas de "síntese neural" que produzem resultados de áudio no domínio da forma de onda, tornou-se possível gerar dezenas de horas de música. Embora, no momento, a síntese não alcance um nível profissional, notamos que a indústria do áudio publiciza esses sistemas como "assistentes composicionais", em especial devido seu potencial em fornecer muitas variações de ideias musicais presentes nas amostras sonoras originais. Sobre esse aspecto, a partir das pesquisas de Carr e Zukowski, ficou claro que é tedioso para um artista explorar a gama musical completa do *output* dessas redes para eventualmente selecionar material musical relevante na aplicação em contextos musicais reais. Nesse sentido, estratégias de curadoria e técnicas de visualização desses dados, em especial utilizando tecnologias de *Self Organizing Maps*, se tornam cada vez mais necessárias (CARR e ZUKOWSKI, 2019; MUNTREF, 2020). Desdobramentos nessa área podem ser observados em iniciativas como o já mencionado programa *AudioStellar* (MUTREF, 2020) – que apresenta grande auxílio na visualização e manipulação de grandes amostragens de áudio digital.

Quanto aos próximos passos da presente pesquisa, ressaltamos a necessidade de experimentos continuados em situações artísticas reais, junto de implementações em sistemas computacionais de maior acesso aos artistas brasileiros. Esperamos que com essas medidas possamos melhor avaliar a escalabilidade expressiva dessas tecnologias no contexto nacional. Pontuamos também que é de nosso interesse investigar alternativas técnicas que possam otimizar os métodos de síntese e aprimorar a qualidade sonora de amostragens futuras.

Por fim, concluímos que as áreas de pesquisa em ciência da informação, aprendizado de máquinas e artes aplicadas cada vez mais se sobrepõem e coabitam. Nesse sentido, não podemos ignorar que questões filosóficas e políticas devam ser incluídas no desenvolvimento técnico de IA, visto que essas áreas problematizam de forma construtiva o uso ético de aprendizado de máquinas no campo estético e social contemporâneo.

REFERÊNCIAS

Aiva Technologies (2020). *Aiva*. Disponível em <https://www.aiva.ai/>. Acesso em: 25 de set. 2020.

Amoore, Louise (2020). *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Londres: Duke University Press.

Arik, S. O. et al (2017). *Deep Voice: Real-time Neural Text-to-Speech*. Disponível em: <arXiv:1702.07825>. Acesso em: 25 de set. 2020.

Broussard, Meredith (2018). *Artificial Unintelligence: How Computers Misunderstand the World.* Cambridge: The MIT Press.

Caillon, Antonie e ESLING, Philippe. *Streamable Neural Audio Synthesis with Non-Causal Convolution*. Disponível em : <https://arxiv.org/pdf/2204.07064.pdf>. Acesso em: 15 de junho de 2023.

Carr, Cj e Zukowski, Zack (2017). *Generating Black Metal and Math Rock: Beyond Bach, Beethoven and Beatles*. 31st Conference on Neural Information Processing System, NIPS. Disponível em: <https://arxiv.org/abs/1811.06633>. Acesso em: 27 de set. 2020.

_____(2018). *Generating Albums with SampleRNN to Imitate Metal, Rock and Punk Bands*. MUME. Disponível em: <https://arxiv.org/abs/1811.06633>. Acesso em: 27 de maio de 2021.

_____(2019). *Curating Generative Raw Audio Music with D.O.M.E.* MILC. Disponível em: <http://ceur-ws.org/Vol-2327/IUI19WS-MILC-3.pdf>. Acesso em: 27 de mai. 2021.

Dadabots (2019). *Relentless Doppelganger*. Dadabots YouTube Channel. Disponível em: < https://www.youtube.com/watch?v=MwtVkPKx3RA>. Acessado em 28 de ago. de 2021.

_____(2021). *Music Page*. Dadabots. Disponível em: <https://dadabots.com/music.php>. Acessado em 28 de ago. de 2021.

Dhariwal, Prafulla, et. al (2020). *Jukebox: A Generative Model of Music.* OpenAI. Disponível em: <https://openai.com/blog/jukebox/>. Acesso em 28 de setembro de 2020.

Dvs Sound (2017). *Hybrid Vehicle with a LOM Elektrosluch 3+-HQ reversed 001*. Dvs Sound YouTube Channel. Disponível em: <https://www.youtube.com/watch?v=kz0eL_RmCQg&t=83s>. Acesso em: 25 de set. 2020.

Eck, Douglas (2016). *Welcome do Magenta!* Google AI. Disponível em <https://magenta.tensorflow.org/blog/2016/06/01/welcome-to-magenta/>. Acesso em: 25 de set. 2020.

Engel, Jesse, et al (2019). *GANSynth: Adversarial Neural Audio Synthesis.* Google AI. Disponível em: <https://openreview.net/forum?id=H1xQVn09FX>. Acesso em: 25 de set. 2020.

Engel, Jesse e Resnick, Cinjon, et al (2017). *Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders.* Google Research. Disponível em <https://research.google/pubs/pub46119/>. Acesso em: 25 de set. 2020.

Eubanks, Virginia (2018). *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor.* Nova Iorque: St. Martin's Press.

Facebook (2021). *Pytorch*. Disponível em: <https://pytorch.org.>. Acesso: 13 de ago. 2021.

Fedden, Leon (2017). *Comparative Audio Analysis with WaveNet, MFCCs, UMAP, t-SNE and PCA.* Medium. Disponível em: <https://medium.com/@LeonFedden/comparative-audio-analysis-with-wavenet-mfccs-umap-t-sne-and-pca-cb8237bfce2f>. Acesso em: 25 de jun. 2021.

Google a. (2021). *TensorFlow 2*. Disponível em: <https://tensorflow.org>. Acesso: 13 de ago. 2021.

Google b (2021). *Deep Dream Generator.* Google. Disponível em: <https://deepdreamgenerator.com>. Acesso em: 27 de ago. 2021.

Graves, A (2013). *Generating Sequences with Recurrent Neural Networks.* Disponível em: <https://arxiv.org/abs/1308.0850>. Acesso em: 27 de maio de 2021.

Gray, Mary L. e Suri, Siddharth (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass.* Nova Iorque: Houghton Mifflin Harcourt Publishing Company, 2019.

Herdon, Holly (2021). *Holly Plus*. Never Heard Before Sound. Disponível em: <https://holly.plus>. Acesso em: 27 de ago. 2021.

Hiner, Karl (2019). *Generating Music with WaveNet and SampleRNN*. Disponível em: <https://karlhiner.com/music_generation/wavenet_and_samplernn/>. Acesso em: 27 de ago. 2021.

Hochreitter, S. e Schmidhuber, J (1997). *Long Short-Term Memory.* Neural computation, 9(8): 17351780.

Huang, Cheng-Zhi, et al (2018). *Music Transformer: Generating Music with Long-Term Structure*. Cornell University. Disponível em: <https://arxiv.org/abs/1809.04281>. Acesso em: 25 de set. 2020.

Lemos, Gabriel Francisco (2016). *Binah.* Disponível em <https://vimeo.com/358627864>. Acesso: 25 de ago. 2021.

Kalchbrenner, N. et al (2018). *Efficient Neural Audio Synthesis.* Disponível em: <arXiv:1802.08435>. Acesso em: 27 de maio de 2021.

Karpatchy, A (2015). *The Unreasonable Effectiveness of Recurrent Neural Networks.* Disponível em: <http://karpathy.github.io/2015/05/21/rnn-effectiveness/>. Acesso em: 27 de maio de 2021.

Maaten, Laurens van der; Hinton, Geoffrey (2008). *Visualizing Data t-SNE*. Journal of Machine Learning Research, Volume 9, p. 2579-2605.

Mehri, Soroush, Kumar, Kundan, Gulrajani, Ishaan, Kumar, Rithesh, Jain, Shubham, Sotelo, Jose, Courville, Aaron C., and Bengio, Yoshua (2016). *Samplernn: An unconditional end-to-end neural audio generation model.* CoRR, abs/1612.07837. Disponível em: <http://arxiv.org/abs/1612.07837>. Acessado em: 25 de set. 2020.

Melen, Christopher (2020). *A Short History of Neural Synthesis.* Manchester: Research Centres at the RNCM. Disponível em: <https://www.rncm.ac.uk/research/research-centres-rncm/prism/prism-blog/a-short-history-of-neural-synthesis/>. Acesso: 13 de ago. 2021.

Muntref (2020). *AudioStellar.* Muntref Centro de Arte y Ciencia. Disponível em: <https://audiostellar.xyz>. Acesso: 13 de ago. 2021.

Norvig, Peter e Russell, Stuart (2021). *Artificial Intelligence a Modern Approach.* 4a Edição. Pearson Editions.

Perceptron (2011). *Redes Neurais Artificiais Blogspot*. Disponível em: <http://redesneuraisartificiais.blogspot.com/2011/06/perceptron-uma-breve-explicacao.html**>.** Acesso: 13 de ago. 2021.

Salem, Sam (2021). *Prism-SampleRNN*. Github. Disponível em: <https://github.com/rncm-prism/prism-samplernn>. Acesso em: 28 de maio de 2021.

Schubert, Alexander (2021). *Switching Worlds*. Vorlke-Verlag. Disponível em: <https://www.wolke-verlag.de/wp-content/uploads/2021/02/SwitchingWorlds_DIGITAL_englisch_210222.pdf>. Acesso em: 19 de fev. 2021.

Schultz, D. V. (2021). *StyleGAN2-ADA*. GitHub. Disponível em: <https://github.com/dvschultz/stylegan2-ada>. Acesso em: 27 de ago. 2021.

Steyerl, Hito (2017). *Duty Free Art: Art in the Age of Planetary Civil War*. Nova Iorque: Verso.

Van Den Oord, Aäron e et al (2016). *Wavenet: A Generative Model for Raw Audio*. CoRR, abs/1609.03499. Disponível em: <http://arxiv.org/abs/1609.03499>. Acesso em: 19 de set. 2019.

Veen, Fjodor Van (2016). *The Neural Network Zoo.* The Asimov Institute. Disponível em: <https://www.asimovinstitute.org/neural-network-zoo/>. Acesso em 25 de jun. 2021.

Vickers, Ben e Allado Mcdowell, K. (orgs.) (2021). *Atlas of Anomalous AI*. Londres: Ignota Books.

Wikipedia (2021). *Linear Regression*. Disponível em: <https://en.wikipedia.org/wiki/Linear_regression>. Acesso em: 19 de fev. 2021.

Zhang, Jin (2008). *Visualization for Information Retrieval*. Berlim: Springer-Verlag.