GenIR: Génération de réponse impulsionnelle neuronale guidée par le texte pour une réverbération créative

Michael Zajner
Music Technology
555 Sherbrooke St. W.
Montreal, Quebec, Canada
michael.zajner@mail.mcgill.ca

Arthur Schick
Musiques numériques
200 Av. Vincent-D'Indy, Outremont
Montreal, Quebec, Canada
arthur.schick@umontreal.ca

ABSTRACT
La réverbération à convolution est aujourd'hui devenue un élément essentiel de la production musicale et de la conception sonore. C'est pourquoi la variété et la qualité des réponse impulsionnelles audio disponibles sur internet s'est extrêmement enrichie au cours du temps. Parmi ces données sonores, une grande quantité sont issues de lieux réels enregistrés, mais à quoi ressemble l'espace entre ces lieux que nous connaissons bien ? Nous présentons GenIR, un nouveau système interactif qui utilise l'intelligence artificielle pour générer des réponses impulsionnelles audio à partir de descriptions textuelles et de paramètres modifiables. En s'appuyant sur les avancées récentes en matière de génération audio neuronale et de modèles de diffusion latente, GenIR permet aux utilisateurs de générer des espaces réverbérants uniques et de les appliquer dans le traitement de la réverbération à convolution.

Nous décrivons l'architecture du système, qui intègre une interface utilisateur graphique avec un modèle génératif basé sur le cloud accessible via une API Gradio, enveloppée dans un format VST3 pour la production audio. Nous discutons également des applications créatives de cette technologie, de la création de paysages sonores immersifs à l'inspiration de nouvelles textures musicales, et nous réfléchissons à l'évolution de l'agentivité entre les musiciens humains et les outils d'IA. Les évaluations préliminaires indiquent que GenIR est à la fois utilisable et inspirant pour les musiciens, tout en soulignant les domaines à améliorer. Ce travail élargit les possibilités de conception de la réverbération, mêlant l'innovation technique à l'exploration artistique.

Author Keywords
Génération de réponses impulsionnelles; réverbération à convolution; audio génératif; text-to-audio; IA créative ; interaction humain-machine.

1. INTRODUCTION
La convolution linéaire est une méthode de traitement bien connue des ingénieurs du son et des développeurs de plugins audios qui permet assez généralement de reproduire artificiellement la réponse acoustique d'un système visé (lieux, haut-parleur, microphone...). C'est un processus assez direct qui se base sur le fait qu'en connaissant la RI de ce système, on peut reproduire l'équation qui le caractérise et l'appliquer à toute autre source sonore [1]. Pour chaque échantillon de sortie, nous calculons alors une somme pondérée des échantillons d'entrée, où les poids sont précisément les valeurs de la réponse impulsionnelle. Cette méthode est utilisée dans divers effets de traitement audio tels que les filtres récursifs, les pédales d'effets, certains types de distorsion, mais surtout, dans l'effet au cœur du plugin que nous avons développé : la réverbération à convolution.

La réverbération à convolution, apparue dans les années 2000, a ouvert le champ des possibles en permettant donc, comme nous l'avons expliqué, de recréer n'importe quel système en disposant seulement de sa réponse impulsionnelle. Mais le véritable nouveau paradigme que cette technique de traitement a apporté pour les utilisateurs, c'est le fait de reproduire mathématiquement et très précisément des espaces acoustiques existants, en ne se servant non pas de calculs paramétriques et d'approximations théoriques, mais bien d'enregistrements réels de ces espaces. Ainsi, toute personne ayant une affection particulière pour le son d'un espace, ou voulant reproduire cet espace exact artificiellement, est devenue en mesure de le faire avec un simple fichier RI et un algorithme de convolution. Cette accessibilité nouvelle a menée à une création et un enrichissement progressif des données de réponses impulsionnelles, qui sont aujourd'hui très fournies et variées (nous pourrions cites les packs EchoThief, BOOM Library Waves ou Altiverb).

Nous en arrivons donc à cet état de fait, ou nous avons à disposition un grand nombre de données audios très contextuelles et auditivement proches, qui décrivent une multitude de systèmes très variés. C'est en réalisant cela que nous avons eu une idée assez simple en apparence : pourquoi ne pas entrainer un modèle de génération audio sur la base de ces données afin de potentiellement pouvoir créer une quantité infinie de réponses impulsionnelles entièrement modelables ? En particulier, la possibilité d'obtenir des réponses impulsionnelles situées à l'intersection de deux espaces acoustiques réels nous semblait éminemment prometteuse et novatrice. En combinant l'exactitude de la convolution et l'approximation de la génération par apprentissage machine pour créer des lieux imaginaires basés sur des lieux réels très précis, nous pensions pouvoir arriver à un entre deux intéressant.

2. REVUE DE LA LITTÉRATURE
2.1 Génération de réponses impulsionnelles basée sur l'IA
L'avènement de la réverbération à convolution a permis d'utiliser n'importe quel IR enregistré à des fins de réalisme, mais la création de certaines IR associées à des lieux difficiles d'accès ou l'élargissement des bases de données disponibles ont été des motivations pour certains chercheurs.

Singh et al. ont introduit une méthode « image-to-reverb » qui synthétise directement une RI à partir d'une photographie d'un environnement [3]. Leur système, Image2Reverb, utilise un réseau neuronal pour faire correspondre des caractéristiques visuelles à une réponse impulsionnelle audio, ce qui permet d'obtenir des réverbérations plausibles même pour des espaces imaginaires ou inaccessibles. Nous avons utilisé un processus plus ou moins similaire en nous basant sur des images pour extraire, à l'aide de PIL, des descripteurs audios propres aux lieux où les IRs que nous avons utilisé pour notre base de données avaient été enregistrées. Nous en parlerons plus en détail dans le chapitre 4.

Une autre approche, IR-GAN, utilise un réseau génératif pour créer des réponses impulsionnelles synthétiques (RIR) après avoir appris les paramètres acoustiques à partir de mesures réelles [4]. Ces réponses générées sont ensuite utilisées pour augmenter les données

d'entraînement pour la reconnaissance vocale en champ lointain, ce qui permet d'améliorer la précision de la reconnaissance vocale pour les lieux où les véritables RIR étaient rares. D'autres travaux ont cherché à accroître la contrôlabilité des RIR générés par l'IA - par exemple, en conditionnant les réseaux neuronaux à des paramètres explicites de la pièce (tels que les dimensions ou les matériaux) pour guider les caractéristiques de la réverbération.

2.2 Synthèse et outils audio neuronaux

Parallèlement, le domaine plus large de la génération audio neuronale a mûri. Des modèles autorégressifs comme WaveNet ont démontré que les réseaux neuronaux peuvent synthétiser des formes d'ondes audio réalistes à partir de distributions apprises, bien qu'avec un coût de calcul élevé. Plus récemment, les modèles génératifs latents ont obtenu des résultats remarquables. Le modèle RAVE de Caillon et Esling, par exemple, utilise un autoencodeur variationnel pour produire rapidement un son de “haute fidélité”, atteignant même une sortie de 48 kHz à des vitesses supérieures au temps réel [5].

Les modèles de diffusion sont également apparus comme un paradigme puissant. Par exemple, AudioLDM est un cadre de diffusion latente qui peut synthétiser des effets sonores généraux à partir de descriptions textuelles [6]. Dans le domaine musical, des modèles à grande échelle tels que MusicLM peuvent générer une musique cohérente de plusieurs minutes à partir d'une entrée textuelle, ce qui témoigne de l'évolutivité de ces méthodes [7]. D'autres travaux ont intégré les réseaux neuronaux dans des outils de production audio. Des boîtes à outils interactives d'apprentissage et mapping automatique (par exemple, Wekinator) ont permis aux artistes de former des correspondances personnalisées entre les gestes et les sons, préfigurant les instruments actuels pilotés par l'IA [8].

3. APERÇU DU SYSTÈME

3.1 Les motivations

Un grand nombre de modèles de générations intelligente de fichiers audios sont entrainés sur un très large éventail de données non spécialisées et sont par conséquent assez mauvais à reproduire les IRs en particulier. En plus de cela, très peu de plugin fournissent actuellement un service de génération par machine learning intégré, et c’est à notre avis une des causes qui fait que l’IA est encore assez peu présente dans les pipelines de production audio. GenIR permet de combler ce manque en offrant un plugin de convolution par IA tout en un.

3.2 Description du projet

GenIR se compose de deux éléments principaux : une interface utilisateur graphique (GUI) avec module de convolution côté client et un moteur audio génératif côté serveur. Le produit côté client est une application VST3 (construite avec Qt pour la gestion réseau et JUCE pour le traitement audio et l’interface graphique) qui permet à l'utilisateur de saisir une description textuelle et d'ajuster les paramètres de génération, tandis que le coté serveur héberge le modèle d'IA et effectue la génération audio. Le client et le serveur communiquent via leur connexion internet en utilisant de simples requêtes HTTP (via l'API Gradio sur le serveur). Cette conception permet au modèle lourd en termes de calcul de fonctionner sur une machine équipée d'un GPU (qui pourrait être un serveur en nuage ou un poste de travail local), tout en conservant la légèreté du client. Un des objectifs principaux était de rendre le plugin le plus accessible possible, ne nécessitant qu’un fichier vst ou exe pour fonctionner. Nous voulions nous adresser à tous les utilisateurs de plugins standards, groupe dont nous faisions nous même partis.

3.3 Interaction avec l'utilisateur

Lors du lancement de GenIR, deux onglets différents sont disponibles. Le premier concerne la génération de réponse impulsionnelles, tandis que le deuxième contient les contrôles relatifs à la convolution de cette réponse impulsionnelle. Dans le premier onglet, l'utilisateur se connecte au serveur (c’est une option temporaire qui sera retirée pour la version de production avec un serveur permanent). Il peut ensuite taper une invite de mots-clés décrivant l'espace acoustique souhaité ou les qualités de la réverbération (par exemple, « petite chapelle pierre piliers écho »). Des commandes permettent de définir la durée cible de la RI (de 1 à 20 secondes), le nombre d'étapes de diffusion, l'échelle de guidage et une graine aléatoire (avec une option de randomisation pour chaque génération). L'interface propose également une liste d'exemples d'invites et de mots-clés pour inspirer l'exploration. Après avoir configuré ces différents paramètres, l'utilisateur clique sur Générer, et la demande est envoyée au serveur. Une fois que le serveur renvoie la RI générée (sous forme de fichier audio), celle-ci est enregistrée automatiquement comme fichier temporaire et le vst permet à l'utilisateur de l'écouter immédiatement ou de la sauvegarder.

En passant au deuxième onglet, l'utilisateur peut alors accéder à divers contrôles sur la convolution. Il s'agit notamment 1) d'un gain d'entrée pour atténuer le signal d'entrée ; 2) d'un gain de sortie pour contrôler le signal de sortie ; 3) d'un mélange sec/humide pour équilibrer le signal d'entrée brut avec le signal convolué ; 4) d'un paramètre d'amortissement pour contrôler la fréquence de coupure d'un filtre passe-bas de type IIR ajouté au signal de rétroaction. En plus, un menu déroulant permet à l'utilisateur de sélectionner et de charger des RI préfabriqués ainsi que de charger ses propres RI à partir d'un explorateur de fichiers.

3.4 Fonctionnement côté serveur

Le serveur héberge un modèle de génération text-to-audio basé sur l’architecture Tangoflux et entrainé sur des IRs de la librairie OpenAI étiquetées [9]. Nous utilisons ensuite Gradio pour exposer le modèle sous la forme d'une simple API web. Lorsque le client émet une demande de génération, elle appelle une fonction de prédiction Gradio avec le texte et les paramètres d'entrée, qui sont tous convertis pour être compris par l’API python. Le modèle de diffusion du serveur synthétise alors la réponse impulsionnelle et la stocke dans un fichier WAV temporaire sur la page internet, lequel est récupéré par l’application et stocké dans les fichiers temporaires locaux du client. Cette approche client-serveur, contrairement à une approche de scripts exportés avec TorchScript, permet d’une part d’alléger le plugin qui serait quasiment inutilisable par un utilisateur lambda autrement, et d’autre part d’offrir une réelle exploration de toutes les combinaisons possibles, laquelle aurait dû être beaucoup plus limitée si l’on avait voulu utiliser TorchScript.

4. MÉTHODOLOGIE

Tangoflux, le modèle génératif au cœur de GenIR, est basé sur une architecture de diffusion latente pour l'audio. C’est un VAE (Variational auto-encoder) text-to audio opérant dans un espace de représentation compressé, similaire à AudioLDM [6] [9]. Nous avons choisi un VAE car nous voulions pouvoir manipuler l’espace latent dans de futures versions et un modèle text to audio car nous trouvions que les descriptions physiques d’un lieu marchaient bien pour se le représenter, dans l’interaction avec l’utilisateur. TangoFlux, modèle assez récent et relativement régulièrement mis à jour, proposait en plus un traitement beaucoup plus rapide des demandes de génération que les autres modèles actuellement disponibles (Stable Audio, AudioLDM2). Une grosse partie du travail a été de mettre en place un bon étiquetage, lequel est nécessaire au bon fonctionnement du modèle. Pour ce faire, nous avons mis au point le script ajout_descripteurs.py, qui est disponible en annexe, et qui utilise

une fonctionnalité de PIL de description d'image entrainée elle-même par machine learning. Ce script nous a permis d'obtenir certains descripteurs sonores basé sur les images des lieux relatifs aux IRs, mais un gros travail de correction et d'augmentation a tout de même dû être fait à la main.
L'entrainement effectué a été assez court puisque la base de données est très réduite (version de test), mais le modèle résultant permet quand même des variations intéressantes, en particulier lorsque l'on joue avec les différents paramètres. Les fichiers résultants sont des fichiers WAV 44100 Hz 24 bits Stéréos qui durent entre 1 et 20 secondes. L'interface graphique envoie des requêtes via HTTP dans un thread d'arrière-plan pour éviter de bloquer l'interface. Le serveur utilise PyTorch pour générer de l'audio, sur un GPU NVIDIA avec cuda.

En pratique, lorsque l'utilisateur déclenche la génération, le serveur traite l'invite et renvoie un fichier audio en quelques secondes. L'utilisateur peut alors écouter la RI ou la reverb et, s'il n'est pas satisfaisant, modifier la description ou les paramètres et réessayer. Dans notre implémentation actuelle, une RI typique de 5 secondes (avec 50 étapes de diffusion) prend environ 3 secondes à générer sur le GPU du serveur. Les durées plus longues s'échelonnent de manière à peu près linéaire dans le temps. La surcharge de communication (envoi d'un court texte et réception d'un petit fichier audio) est négligeable par rapport au temps de génération. Ainsi, l'utilisateur ne subit qu'une courte attente et peut ensuite immédiatement écouter le résultat.

5. CONCLUSION

L'introduction d'un agent d'intelligence artificielle dans le processus de conception de la réverbération invite à réfléchir à l'évolution de la relation entre les musiciens et la technologie. Avec GenIR, un musicien n'est plus limité à la sélection de préréglages ou à la bonne volonté des preneurs de son spécialisés dans les réponses impulsionnelles ; au lieu de cela, il décrit une idée avec des mots et reçoit un résultat façonné par un modèle complexe entrainé sur des données d'espaces réels. Cela soulève des questions d'agentivité et de coauteur créatif. Dans la pratique, l'agence est partagée : l'humain fournit l'intention sémantique et la machine fournit la réalisation détaillée. Cette dynamique s'aligne sur des observations antérieures selon lesquelles l'apprentissage automatique peut servir de partenaire ou de matériau dans la création d'instruments plutôt que d'outil déterministe [10].

Il est important de noter que l'utilisation de GenIR est un acte d'exploration. Le système peut surprendre l'utilisateur avec des résultats qui s'écartent de ses attentes - par exemple, une invite pour une « grotte forestière » produit une RI avec des tonalités résonnantes imprévues. Dans certains cas, les bizarreries imprévisibles du modèle peuvent produire de nouvelles textures - un cas de découverte d'« accidents heureux » créatifs dans la production d'une machine, comme l'ont observé d'autres praticiens [11]. En réalité, la quasi-totalité des réponse impulsionnelles créées sont assez peu naturelles, même avec des espaces décrits complètement réalistes, et cela agit bien comme signature de l'utilisation d'intelligence artificielle. Peut-être qu'au final c'est même plutôt cela l'intérêt premier du plugiciel : pouvoir créer et entendre directement les effets d'une réponse impulsionnelle marquée de ce son caractéristique de la génération par IA. Ces réponses deviennent par le fait même uniques en leur genre et c'est peut-être là l'essence et la singularité d'un plugin comme GenIR. Le musicien s'engage donc dans un dialogue avec l'outil, en affinant les messages ou en acceptant et intégrant les résultats inattendus dans une boucle de rétroaction.

Des études d'utilisateurs sur l'attitude des musiciens à l'égard des outils d'IA ont révélé que de nombreux praticiens considèrent ces technologies comme des extensions de leur art, et non comme des menaces ou des substituts [10]. GenIR est conçu dans cet esprit : il élargit la palette de réverbérations du sound designer, mais c'est l'humain qui en contrôle la sélection et l'utilisation. L'outil ne décide pas de manière autonome de ce qui constitue une « bonne » réverbération ; il offre des possibilités parmi lesquelles l'artiste fait son choix. L'implication philosophique est que l'intention créative reste chez l'utilisateur humain, même si les moyens de réalisation deviennent de plus en plus autonomes. En ce sens, GenIR peut être considéré comme un instrument ou un collaborateur intelligent qui suit l'utilisateur. Sa présence dans le processus créatif encourage le musicien à considérer la réverbération à un niveau conceptuel (imaginer des espaces) et à adopter une approche plus intentionnelle et imaginative de la conception sonore.

D'un point de vue plus large, les systèmes de co-création comme GenIR remettent en question les flux de travail traditionnels : en même temps qu'ils règlent les boutons, les créateurs s'engagent dans une interaction descriptive. Nous avons cherché à concevoir GenIR avec un certain degré de contrôle et d'interprétabilité utilisateur (par exemple, en limitant l'automatisation et en tenant l'utilisateur informé à chaque génération), mais l'expérience du plugin est essentiellement basée sur un processus d'essai erreur, dû au fonctionnement non-explicite du modèle de génération qu'il utilise. GenIR démontre que l'intégration de l'IA peut agir comme un puissant outil de réverbération améliorée, permettant aux artistes d'explorer de nouvelles possibilités acoustiques tout en préservant leur volonté artistique.

6. RÉFLEXION SUR LE PLUGICIEL

Pour l'avenir, notre feuille de route est centrée sur l'élargissement de l'accessibilité et du contrôle, ainsi que l'enrichissement du modèle et la stabilisation du serveur. Nous avons déjà développé un plugin VST pour Windows uniquement, qui intègre des invites textuelles et des paramètres de génération directement dans les stations de travail audionumériques ; le support Mac suivra. Un des champs d'actions qui pourrait être intéressant à investiguer est la génération d'IRs binauraux et le conditionnement de cette génération à des cibles acoustiques explicites (par exemple, RT60, densité de réflexion précoce), en mélangeant l'intuitivité du texte avec la précision des contrôles de réverbération traditionnels. Un autre gros intérêt encore inexploité du plugiciel pourrait être la navigation de l'espace latent explicite, via des contrôles graphiques, mais cela impliquerait probablement un traitement génératif en live des réponses impulsionnelles, pour l'instant impossible à mettre en place.
Grâce à une utilisation pratique, GenIR s'est avéré inspirer de nouveaux flux de travail créatifs : les concepteurs décrivent des espaces imaginés, reçoivent une IR personnalisée, puis l'auditionnent et la traitent immédiatement via le module de convolution. Cette pratique réflexive - entre autres, l'itération d'invites à la recherche de résonances - positionne GenIR comme un partenaire créatif plutôt que comme un navigateur préétabli.

En étendant la prise en charge des plateformes, en améliorant le contrôle et en optimisant les performances avec GenIR, nous espérons parvenir à faire le lien faire le lien entre l'imagination et la réalité acoustique, et à permettant aux concepteurs sonores et aux musiciens de sculpter des espaces sonores inexplorés grâce à la créativité des systèmes intelligents.

7. REFERENCES

[1] Yee-King, MJ, 2024, Build AI-Enhanced Audio Plugins with C++, Focal Press, ISBN : 9781032430423

[2] Vesa Välimäki, Julian D. Parker, Lauri Savioja, Julius O. Smith, and John S. Abel. 2012. *Fifty years of artificial reverberation*. IEEE Transactions on Audio, Speech, and Language Processing 20(5): 1421–1448.

[3] Nikhil Singh, Jeff Mentch, Jerry Ng, Matthew Beveridge, and Iddo Drori. 2021. *Image2Reverb: Cross-Modal*

Reverb Impulse Response Synthesis. In Proceedings of ICCV 2021. (Also available as arXiv:2103.14201).

[4] Anton Ratnarajah, Zhenyu Tang, and Dinesh Manocha. 2021. *IR-GAN: Room Impulse Response Generator for Far-field Speech Recognition*. arXiv:2010.13219 [cs.SD].

[5] Antoine Caillon and Philippe Esling. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. arXiv:2111.05011 [cs.SD].

[6] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. 2023. *AudioLDM: Text-to-Audio Generation with Latent Diffusion Models*. In Proceedings of ICML 2023.

[7] Andrea Agostinelli, Timo I. Haber, et al. 2023. *MusicLM: Generating Music from Text*. arXiv:2301.11325 [cs.SD].

[8] Rebecca Fiebrink and Laetitia Sonami. 2020. *Reflections on Eight Years of Instrument Creation with Machine Learning*. In Proceedings of the International Conference on New Interfaces for Musical Expression (NIME 2020).

[9] Hung C.-Y., Majumder N., Kong Z., Mehrish A., Valle R., Catanzaro B., and Poria S., "TANGOFLUX: Super Fast and Faithful Text-to-Audio Generation with Flow Matching and CLAP-Ranked Preference Optimization," *arXiv* preprint arXiv:2412.21037 [cs.SD], Apr. 2025.

[10] Bob L. T. Sturm and Oded Ben-Tal. 2021. *Machine Learning for Folk Music Generation: A Critical Reflection*. In Music AI (eds. T. Collins et al.), Springer, pp. 443–450.

[11] Shelly Knotts and Nick Collins. 2021. *Musicians' Attitudes Toward Creative AI: A Survey*. In Music AI (eds. T. Collins et al.), Springer, pp. 843–851.