McGill
Schulich School of Music of McGill University
555 Sherbrooke Street West
Montreal, Quebec H3A 1E3

Schulich School of Music
École de musique Schulich
École de Musique Schulich
de l'Université McGill
555, rue Sherbrooke Ouest
Montréal (Québec) H3A 1E3

(514) 398-4535
(514) 398-8061
www.mcgill.ca/music

À : M. Habib Mehrez, Direction de l'École doctoral 130
Sorbonne Université,
21 rue de l'école de médecine
75006 Paris.

De : Philippe Depalle, rapporteur, Université McGill, Montréal, Québec, Canada

Date : 19 novembre 2023

Objet : Rapport sur la thèse d'Antoine Lavault
Niveau : Doctorat en Sciences et technologies de l'information et de la communication
Spécialité : Sciences et technologie de la musique et du son
Titre : Réseaux antagonistes génératifs pour la synthèse et le contrôle des sons de batterie (*Generative Adversarial Networks for Synthesis and Control of Drum Sounds*).

Avis : Favorable pour une soutenance en l'état

Les travaux de recherche présentés par Antoine Lavault s'inscrivent dans le cadre général de la synthèse des sons à des fins musicales avec contrôle de jeu de haut-niveau. Plus spécifiquement son projet est consacré à la synthèse en temps-réel de sons de batterie contrôlée par des descripteurs de timbre et de dynamique reposant sur des techniques d'apprentissage profond. Antoine Lavault part du constat que l'utilisation, bien que récente, des techniques d'apprentissage profond et plus particulièrement des réseaux antagonistes génératifs (GAN) pour la synthèse de signaux a atteint un niveau de maturité et de qualité certains. La thèse qu'il défend est que les GANs sont intrinsèquement aptes à rendre compte non seulement de la complexité des signaux sonores (synthèse), mais également des liens entre leurs qualités (timbre) et les descripteurs qui ont été développé pour en rendre compte. Ce lien qui possède des atouts opérationnels font des GANs des candidats de choix pour le contrôle de la synthèse. Le candidat propose une suite de solutions et d'expérimentations afférentes qui aboutit à un dispositif hybride de GANs et de traitement de signal.

Les contributions du candidat sont avérées : elles résident dans le choix et le raffinement des techniques et algorithmes d'apprentissage afin de les adapter aux spécificités des sons percussifs, ce qui aboutit au développement du système de synthèse que le candidat appelle *StyleWaveGAN* et doté de diverses variations structurelles. Elles concernent la constitution d'une base de données extensible de sons de batterie annotée avec information de vélocité, dénommée ApeiraDrums, d'une approche de type réseaux de neurones pour le contrôle du son de synthèse par des caractéristiques du timbre, et de l'intégration d'une stratégie de contrôle de la vélocité à partir du profil énergétique des signaux sonores, ainsi que des évaluations objectives et subjectives de ces différents éléments.

Le document remis par le candidat décrit bien les travaux effectués et ce, de manière relativement claire. Il se compose de cinq chapitres et est accompagné d'exemples sonores et de certains codes sources rendus accessibles sur un site Web compagnon. La lecture du document révèle bien le caractère pluridisciplinaire des travaux alliant apprentissage automatique (pour lequel le candidat démontre une grande expertise), synthèse des sons, et techniques d'évaluation tant subjective qu'objective. La perspective globale sur son travail aurait néanmoins gagné en lisibilité par le regroupement de certains éléments disséminés au fil de la thèse et notamment l'ensemble des descriptions de la structure des signaux (tels que l'enveloppe temporelle, le portamento de la structure modale, la description de la partie déterministe par banc d'oscillateurs) qui s'étalent sur deux voire trois chapitres. Aussi les descripteurs de type AudioCommon sont mentionnés dès le chapitre deux, mais décrits au chapitre 4 seulement. Il en est de même pour les bases de données et la description des réseaux de neurones. Enfin, une certaine confusion entre les termes de vélocité (parfois MIDI, parfois mécanique), d'accélération, de puissance et d'énergie obscurcissent le propos. Enfin le document est étayée par de nombreuses références qui reflètent bien l'état de l'art dans lequel le travail du candidat s'insère.

L'introduction, relativement longue, expose de manière claire et succincte les motivations et les objectifs du problème à résoudre, à savoir la réalisation d'un synthétiseur de sons de batterie de haute qualité, temps-réel avec contrôle du timbre et de la dynamique en situation de jeu, et se termine par la liste des contributions réalisées par le candidat. La présence d'une partie dédiée au contexte (*background*) en forme d'état de l'art sur certains aspects de la physique et des caractéristiques acoustiques des signaux sonores percussifs, de la synthèse sonore, mais également de l'apprentissage profond, du contrôle et de l'évaluation subjective est intéressante mais pourrait faire l'objet d'un chapitre séparé ou bien d'une partie de l'état de l'art (chapitre 2). En ce qui concerne ces derniers éléments, la physique des sons est bien décrite et met en exergue les éléments saillants de la structure des signaux à prendre en compte pour une synthèse de qualité. La partie consacrée à la synthèse sonore aurait pu être plus développée, notamment en mentionnant la synthèse par somme de sinusoïdes amorties qui est en adéquation avec la structure modale de nombreux signaux percussifs. Par ailleurs, l'enveloppe temporelle et la puissance des signaux aurait pu faire l'objet d'une présentation spécifique en raison de l'importance qu'elle revêt dans la structure du synthétiseur *StyleWaveGan* développé par le candidat ainsi que dans le rôle joué pour le contrôle de la dynamique en situation de jeu. Il s'ensuit une bonne description des méthodes d'évaluations subjectives où le candidat note la difficulté à évaluer la qualité dans le cadre la création de sons sans référents existants.

Le chapitre 2 dresse un état de l'art des modèles de réseaux de neurones génératifs, de leurs stratégies de contrôle ainsi que des stratégies d'évaluation. Ce chapitre est très bien étayé et donne une vision précise des travaux récents sur les modèles génératifs bien souvent dédiés à la synthèse d'image ou de parole mais aussi à la synthèse de signaux audio et parfois de signaux percussifs. Les architectures classiques de réseaux de neurones (avec apprentissage profond ou pas) ayant été présentées dans le chapitre 1, le candidat s'attache, dans le chapitre 2, à décrire les modèles génératifs et leurs particularités. Sont ainsi abordés le premier réseau génératif consacré à l'audio (*WaveNet*), appliqué à la parole puis aux sons musicaux, l'approche de type auto-encodage, l'auto-encodage variationnel et son application aux signaux percussifs. La présentation se poursuit avec diverses architectures fondées sur l'approche antagoniste générative (GAN), et leurs applications dans le domaine de l'audio dont plusieurs dédiées aux percussions. Il aborde la technique de traitement de signal numérique différentiable (*DDSP*) qui présente l'avantage

notable d'intégrer des modèles de traitement de signal dans des environnements d'apprentissage profond, et termine par les modèles de diffusion. S'ensuit une description des travaux sur le contrôle des GANs plus spécifiquement tournés vers le contrôle des caractéristiques de timbre. Le chapitre continue avec une description des stratégies d'évaluation des modèles génératifs, et notamment au moyen de la distance de Fréchet et de sa variante audio (FAD) qui sera retenue par le candidat. L'un des buts de ce chapitre étant de motiver le choix d'une architecture spécifique, le candidat termine par une discussion concluant que la structure GAN présente un bon compromis entre qualité sonore, contrôle par paramètres de haut-niveau (timbre et dynamique), et coûts de calcul. Le candidat s'inspirera par ailleurs de certains éléments de l'approche *differentiable* (*DDSP*) qu'il intégrera dans son architecture pour le contrôle des caractéristiques du timbre.

Le chapitre 3 traite d'un élément essentiel et critique pour l'apprentissage profond : la base de données sur laquelle l'apprentissage est effectué. Le candidat s'attache à répertorier les bases de données de sons percussifs existantes et accessibles, procède à des modifications dont il explique la nature, et enfin constitue une nouvelle base de données de sons percussifs, appelée ApeiraDrums. Par souci de cohérence et en adaptation avec la nature du projet, le candidat sélectionne dans les bases de données existantes des notes isolées, sans diaphonie et en champ proche annotées par l'inclusion d'information sur le type de percussion, de son timbre et de sa dynamique. Lorsque la base de données est trop petite, il l'élargit en générant des sons modifiés par variations de l'attaque, du bruit, de la hauteur et de l'enveloppe spectrale à l'aide d'un vocodeur de phase. Enfin le candidat constitue une nouvelle base de données de sons percussifs de haute qualité enregistrés en studio par des professionnels et qui contient des informations fournies par un accéléromètre et un capteur piézoélectrique en vue du contrôle de la dynamique (ApeiraDrums). Le candidat a pris soin de calibrer et de documenter les conditions d'enregistrement afin de pouvoir rajouter des sons dans le futur. Le rapporteur suggère de faire un récapitulatif du matériel employé et des conditions d'enregistrement en annexe de la thèse.

Le chapitre 4 est consacré à la description de *StyleWaveGAN*, le dispositif de synthèse de signaux de percussion que le candidat a conçu et mis en place. Il présente successivement la structure de base du synthétiseur *StyleWaveGAN*, la stratégie de contrôle par descripteurs de timbre, une extension de *StyleWaveGAN* intégrant un banc d'oscillateurs, et le contrôle de la dynamique. C'est indéniablement le chapitre le plus important tant par le nombre de contributions présentées que par la taille puisqu'il représente la moitié du corps principal de la thèse.
Se fondant, sur l'état de l'art, Antoine Lavault imagine une structure de réseau de type GAN avec encodage de style, inspiré de *StyleGAN* mais équipé d'une fonction d'objectif de type WGAN-LP et travaillant sur le signal temporel. Il rajoute au niveau de chaque couche un bruit additionnel, formée temporellement afin d'atténuer le bruit en fin de signal. De la même manière le signal de sortie est affublé d'une enveloppe temporelle, pré-calculée par type de percussion pour d'une part avoir un meilleur contrôle du bruit sur les fins de notes, et d'autre part offrir un meilleur conditionnement énergétique des signaux traités par le réseau. Par ailleurs le candidat précise l'approche utilisée pour générer des types de son spécifiques (tom, cymbale, etc.) en augmentant le vecteur de variables latentes par des étiquettes encodées identifiants ces types de son. Enfin le candidat apporte une amélioration par rapport à l'état de l'art qui consiste à modifier la partie discriminante du synthétiseur par un algorithme hybride entre le *Progressive Growing* et le *Highway* que le candidat appelle *AutoFade*. L'apprentissage est effectué de manière classique sur la base de données ENST augmentée et la qualité audio est comparée à *NeuroDrum*, *WaveGAN*, et *DrumGAN*. L'évaluation montre la qualité supérieure de la méthode du candidat à la fois

objectivement et subjectivement. Il confirme par ailleurs une amélioration due à l'*AutoFade* et une amélioration nette due à l'utilisation des enveloppes temporelles estimées. En ce qui concerne l'évaluation subjective on constate qu'elle est réalisée sur un nombre restreint de participants.

Une fois son réseau de neurones *StyleWaveGAN* mis en place et évalué, le candidat investigue un contrôle du timbre. Il se fonde sur les caractéristiques de timbre de l'*AudioCommons* et sur l'état de l'art existant dans deux études, celle de Ramires et al. et celles de Nistal et al. Contrairement à ces deux approches qui se fondent sur les seules valeurs des descripteurs, le candidat propose d'incorporer les caractéristiques de timbres directement dans la structure d'optimisation de la phase d'apprentissage selon la méthodologie proposé dans le réseau de type *DDSP*. En pratique, le candidat sélectionne trois caractéristiques qui sont la brillance, la profondeur et la chaleur. L'algorithme de calcul des caractéristiques de timbre est intégré sous forme de fonctions différentiables dans le processus d'apprentissage afin de bénéficier des fonctionnalités de différentiation de *TensorFlow* et utilise la norme L1 comme fonction de coût. L'évaluation objective montre que les valeurs de descripteurs de timbre des sons synthétiques sont proches de celles demandées par paramètres de contrôle, et qu'elles respectent par ailleurs le critère d'ordonnancement utilisés dans les précédentes études. Le candidat montre que la relation entre valeurs de paramètres de synthèse et valeurs estimées sur les sons synthétiques sont en relation essentiellement linéaire et que les résultats sont clairement meilleurs que ceux obtenus par l'état de l'art sur cet aspect tant pour un contrôle individuel que pour un contrôle simultané des trois descripteurs choisis. Par ailleurs, le candidat propose une métrique plus élaborée que celle de l'état de l'art pour estimer la qualité du contrôle.

Dans un troisième temps, le candidat propose une version étendue de *StyleWaveGAN* de nature hybride qui introduit un banc d'oscillateurs à amplitude fixe et à fréquence à décroissance exponentielle. C'est une des contributions les plus originales de la thèse qui intègre des contraintes relevant de la physique des percussions tout en s'inspirant encore une fois de l'algorithmique propre à l'approche différentielle de *DDSP*. Les paramètres de la variation de fréquence sont eux-mêmes appris à aide de réseaux de neurones de transfert de style. Diverses contraintes dont un ordonnancement de fréquences sont ajoutées afin d'améliorer le résultat du modèle. Le tout est construit en complément de l'utilisation d'enveloppes temporelles dans la partie signal du modèle et d'une gestion explicite des étiquettes des diverses classes de sons de percussion. Le dispositif s'avère très performant ainsi que le montre une évaluation objective puis subjective. Ainsi l'évaluation objective montre une amélioration sur les signaux possédant une structure modale forte comme les toms tandis que les sons utilisant le redressement d'enveloppe temporelle sont parfois trop bruités.

Ce chapitre se termine sur une quatrième étude qui vise à introduire le contrôle temps-réel de la dynamique des sons en situation de jeu. Il s'agit d'une autre contribution originale qui s'appuie sur l'utilisation simultanée de la base de données ApeiraDrums constituée par le candidat lors de ce projet qui inclut des informations liées à la vélocité, et du capteur commercial *Senstroke*, qui fournit des données de vélocité. Deux approches sont envisagées par le candidat : une reposant sur un réseau de neurones apprenant un descripteur de dynamique de type musical (4 niveaux tels que pianissimo, etc.) et une autre reposant sur un descripteur différentiable fondé sur une mesure de l'énergie du signal. Le candidat montre que la première approche donne de relativement bons résultats lorsque les sons de la base de données sont ceux qui ont servi à entrainer le réseau pour le contrôle de dynamique, mais de moins bon lorsqu'utilisé avec des sons d'une autre base de données. Pour la deuxième approche le candidat montre qu'une relation analytique simple peut être établie entre l'information de vélocité fournie par *Senstroke* et l'énergie exprimée en deciBels. L'intégration à *StyleWaveGAN* selon le principe repris de DDSP montre une réponse

cohérente du synthétiseur aux paramètres de vélocité (de dynamique). Le candidat termine sur l'idée intéressante que l'utilisation de caractéristiques établies par l'expertise et la connaissance, dès lors qu'elles sont disponibles sous forme d'expressions analytiques peuvent être mises à profit dans l'approche *DDSP* et que cela est préférable à l'application d'un réseaux de neurones travaillant sur les seules données

Dans le chapitre 5, le candidat conclut brièvement en récapitulant l'approche utilisée ainsi que les résultats obtenus lors de ses travaux de recherches et ouvre les perspectives pour des travaux futurs.

Vu la qualité, la quantité et le caractère significatif des travaux de recherche et des développements présentés, et vu la bonne qualité du document soumis, je propose d'attribuer le grade de docteur à Monsieur Antoine Lavault et déclare que la thèse peut être soutenue en l'état.

Fait à Montréal, le 19 novembre 2023,

Philippe Depalle, Professeur agrégé
Université McGill
Montréal (Québec), Canada