Écoute musicale artificielle : une approche unifiée IA symbolique et neuronale pour passer d'une écoute locale et réactive à une écoute globale et narrative

Direction : Gérard Assayag ; Encadrement : Nicolas Obin, Jérôme Nika

Contexte et défi

Le « deep learning » (l'apprentissage par réseaux de neurones profonds) a permis des avancées spectaculaires en traitement du signal audio et musical. Les architectures séquence-à-séquence couplées à des stratégies de démêlage de l'information encodée permettent désormais d'apprendre des représentations structurées à partir de productions sonores complexes comme la voix humaine [Obin, 2018; Obin ; 2021 ; Obin, 2022a ; Obin, 2022b] et de les exploiter efficacement dans un mode génératif, pour la synthèse ou la transformation [Obin, 2011]. Ces techniques se montrent également efficaces dans le traitement du signal audio musical, le domaine du MIR (Music Information Retrieval) : depuis le suivi de hauteurs dominantes ou la séparation des pistes instrumentales dans un flux audio polyphonique jusqu'à la transcription musicale automatique, l'analyse de structure musicale, ou la génération de musique. De manière générale, le deep learning permet d'extraire à partir du flux audio musical une représentation latente de l'information acoustique optimisée en fonction des tâches à accomplir (transcription, séparation, etc…) [Boulanger-Lewandowski, 2012, Jansson, 2017, Rachel, 2017]. En particulier, les mécanismes d'attention et les architectures transformers [Vaswani, 2017] permettent de modéliser efficacement l'information de séquences acoustiques sur différents horizons temporels et d'apprendre efficacement des dépendances temporelles à long-terme. Ces systèmes à grande complexité ont été perfectionnés, et permettent désormais une inférence en temps réel [Mohla, 2020; X. Chen, 2021], et ont été appliqués avec succès à des tâches de génération de flux musicaux symboliques [Huang, 2018], ainsi que de classification de flux musicaux acoustiques [Gong, 2021].

La question de l'écoute artificielle est centrale dans les recherches portant sur la co-créativité humain-machine. Les développements récents de ces thématiques ont donné naissance à de nombreux modèles d'agents génératifs capables de s'intégrer dans la production d'un discours musical collectif dans le périmètre d'un « rôle » établi (un exemple trivial étant par exemple « soliste » ou « accompagnateur»). La définition des mécanismes régissant la contribution d'un agent à ce discours associe :

- Une **mémoire musicale** (un modèle génératif construit sur une base de données musicale)

- Un **comportement établi d'action/réaction** (traitement symbolique-symbolique, et conversion symbolique-acoustique) traduisant des spécifications venant d'un utilisateur ou d'un module de perception en

"intentions" : des requêtes au modèle de mémoire musicale dont l'exécution génère les contributions de l'agent au discours musical.

- Une **perception** (conversion acoustique-symbolique) d'un flux musical opéré par une machine d'écoute.

Toutefois, la question de l'évolution de ces mécanismes pilotés par une écoute artificielle n'a principalement été abordée qu'à une échelle locale. Les machines d'écoute permettent en effet d'enrichir la mémoire musicale de l'agent en temps réel [Assayag, 2006] ; la construction locale de d'une séquence musicale guidée par une structure formelle locale pré-établie [Nika, 2017a] ; ou encore de mettre en place une influence « événementielle » [Bonnasse-Gahot, 2016, Borg, 2020] ou la construction d'anticipations à moyen-terme à partir de l'extraction représentations acoustiques simples (hauteurs de note, « timbre », etc.) [Nika, 2017b].

Si les exploitations du deep learning dans les machines d'écoute musicale réactives ont pour l'heure uniquement abordé le champ de l'extraction et de la prédiction locale de structures formelles sous-jacentes [Carsault, 2020], elles semblent toutes indiquées pour relever le défi du passage à l'échelle musicale supérieure : la réalisation d'une écoute à l'échelle macroscopique de la dimension « narrative » d'un flux audio musical. L'association d'un tel modèle d'écoute haut-niveau à des agents génératifs permettrait à ces derniers de déployer des processus de décisions adaptés, et de soutenir un discours musical sur le long-terme. De plus, cette exploration de la narration musicale à grande échelle dans un contexte d'interaction permettrait également de lever le principal verrou des recherches actuelles s'intéressant aux pratiques de composition. Si les modèles les plus récents peuvent générer automatiquement un discours structuré par des esquisses ou scénarios haut-niveau, ces derniers restent des structures formelles locales [Nika, 2021].

<u>Objectif</u>

Ce sujet de thèse transdisciplinaire se situe au carrefour de l'intelligence artificielle, du traitement du signal et de l'information, de la créativité computationnelle avec comme applications la génération automatique de musique et la musicologie computationnelle. Ce projet cherchera à créer une machine d'écoute réactive capable de concilier les deux aspects de l'intelligence artificielle afin de générer un contenu musical cohérent dans le cadre d'une performance musicale temps-réelle. Ce faisant, il permettra l'apprentissage et la reconnaissance des dimensions musicales impliquées dans la conduite de la narration musicale à grande échelle. Cette formalisation trouvera donc également un champ d'application dans les processus de composition en temps différé.

D'une part, l'intelligence artificielle symbolique (comportement symbolique-symbolique et réaction symbolique-acoustique) sous la forme d'agents génératifs agissant à partir de représentations symbolique de la musique et intégrant des

connaissances musicales a priori et des contraintes sur l'évolution du discours musical telles qu'elles sont formulées actuellement; d'autre part, une machine d'écoute permettant d'apprendre (écoute acoustique-symbolique) à un partir d'un flux audio un ensemble de représentations pertinentes pour l'agent génératif, qui pourront être latentes ou explicitement liées à des descripteurs acoustiques, et la construction d'un discours musical structuré sur le long-terme. Par exemple, contrôler la corrélation entre la ligne mélodique et l'accompagnement harmonique, d'identifier la ligne mélodique principale, ou de déterminer les points de rupture dans le discours musical.

Les structures émergentes d'un discours musical construit collectivement constituent un phénomène collectif, et ne se réduisent pas aux structures de jeu individuelles. Afin de représenter l'information musicale de manière pertinente, il semble donc important de développer une écoute basée sur l'information mutuelle entre interprètes, ou couplage, au sein du discours musical [Canonne, 2015]. Par exemple, une piste envisagée repose sur l'utilisation des mécanismes d'attention croisée [C.-F. Chen, 2021] pour apprendre des représentations d'interactions entre les différents acteurs du discours musical.

La thèse sera en particulier focalisée sur la réalisation d'un module d'écoute neuronal conférant aux agents la capacité de repérer quelle est la topologie d'évènements saillants ou pivots, et les dimensions audios pertinentes à écouter dans un signal avec lequel l'agent interagit, ainsi que les modalités de leurs évolutions. Il pourrait s'agir, par exemple, de détecter la transition entre un premier mouvement musical purement rythmique où l'énergie est la dimension dominante, à un second mouvement plus harmonique où l'attention doit donc être portée sur les hauteurs ou autres descripteurs mélodico-harmoniques. Cette compréhension informera enfin leurs processus de décisions, permettant ainsi de générer automatiquement une réaction adaptée au contexte musical issu de l'écoute.

Pour ce faire, les recherches s'articuleront autour des quatre axes suivants :

- L'apprentissage d'une représentation latente ou explicite permettant d'encoder le contexte musicale sur le flux audio mélangé à l'aide d'architectures neuronales capables de capturer les dépendances temporelles à long-terme telles que les réseaux à convolution dilatées ou à mécanismes d'attention classique (Transformer),
- L'apprentissage d'une représentation latente ou explicite permettant de mesurer l'information mutuelle entre les pistes séparées du flux audio (soliste, accompagnement), par l'introduction d'une notion de couplage entre pistes, qui sera implémentée avec, par exemple, les mécanismes d'attention croisée,
- La quantification des concepts de nouveau et de connu en terme de quantité d'information, par exemple avec un formalisme issu de la théorie de

l'information [Dubnov, 2011], ou encore avec la notion d'incertitude dans les réseaux profonds [Abdar, 2021] ,

- La confrontation des prototypes implémentant ces modèles à la validation de musicien.ne.s expert.e.s.

Il s'agira en effet finalement de combiner les différentes représentations dans une architecture commune pour son intégration dans un module d'écoute réactive. Dans un contexte d'interaction, les représentations apprises devront ainsi améliorer la prise de décision d'un agent génératif, en permettant par exemple de contrôler la corrélation entre ce qui est perçu et ce qui est joué par l'agent, ou encore de déterminer les points de rupture dans le discours musical. Enfin, les modèles de structures musicales appris pour construire ce module d'écoute seront également mis à profit dans des processus de composition en temps différé.

Les apprentissages seront réalisés à partir de bases de données disponibles avec les pistes séparées de différents instrumentistes. Pour cela, nous pourrons compter sur la base de données multipistes de duo d'improvisations libres collectées dans le contexte du projet ANR MICA (Clément Canonne, Ircam) ainsi que d'enregistrements multipistes de concerts du festival de Jazz de Montreux (EPFL Meta Media Center / Montreux Jazz Heritage Lab, Resp. Alain Dufaux). La thèse bénéficiera d'un contexte applicatif privilégié au contact des compositeur.ice.s en résidence ainsi que des musicien.ne.s impliqué.e.s dans des projets de recherche et production. Des collaborations artistiques seront menées tout au long de la thèse pour expérimenter les architectures proposées et les valider par un retour expert avec des professionnels en situation de performance et de studio.

Bibliographie

[Abdar, 2021] Abdar, M. et al. (2021) A review of uncertainty quantification in deep learning: Techniques, applications and challenges, *Information Fusion*, *76*, 243-297.

[Assayag, 2006] Assayag, Gérard, et al. "Omax brothers: a dynamic topology of agents for improvisation learning." *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*. 2006.

[Bonnasse-Gahot, 2016] Laurent Bonnasse-Gahot, An update on the SOMax project, Technical report, Ircam, 2016.

[Borg, 2020] Joakim Borg, Somax 2: A Real-time Framework for Human-Machine Improvisation, Technical report, Ircam, 2019, Dynamic Classification Models for Human-Machine Improvisation and, Composition Master report Ircam, Aalborg University , 2020

[Boulanger-Lewandowski, 2012] Boulanger-Lewandowski, N., Bengio, Y., & Vincent, P. (2012). Modeling temporal dependencies in high-dimensional sequences: application to polyphonic music generation and transcription. Paper presented at the Proc. of the 29th International Conference on Machine Learning (ICML).

[Canonne, 2012] Canonne, C. et al. (2012) Cognition and Segmentation In Collective Free Improvisation: An Exploratory Study, in *International Conference on Music Perception and Cognition*.

[Canonne, 2015] Canonne, C. et al. (2015) Individual Decisions and Perceived Form in Collective Free Improvisation, in *Journal of New Music Research*.

[Carsault, 2020] Thibault Carsault. Introduction of musical knowledge and qualitative analysis in chord extraction and prediction tasks with machine learning. Machine Learning [stat.ML]. Sorbonne Universites, UPMC University of Paris 6, 2020.

[X. Chen, 2021] Chen, X. et al. (2021) Developing Real-time Streaming Transformer Transducer for Speech Recognition on Large-scale Dataset, *arXiv preprint arXiv:2010.11395.*

[C.-F. Chen, 2021] Chen, C.-F. et al. (2021) CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification, *arXiv preprint arXiv:2103.14899.*

[Dubnov, 2011], Dubnov et al. "Audio Oracle Analysis of Musical Information Rate", *Proc. of IEEE Semantic Computing Conference*, ICSC2011, Palo Alto, CA, 2011

[Golvet, 2021] Golvet, T. et al. (2021) With, against, or without? Familiarity and copresence increase interactional dissensus and relational plasticity in freely improvising duos, in *Psychology of Aesthetics, Creativity, and the Arts.*

[Gong, 2021] Gong, Y. et al. (2021) AST: Audio Spectrogram Transformer, *arXiv preprint arXiv:2104.01778.*

[Huang, 2018] Huang, C.-Z. et al. (2018) Music Transformer: Generating Music With Long-Term Structure, *arXiv preprint arXiv:1809.04281.*

[Jansson, 2017] Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., & Weyde, T. (2017). Singing voice separation with deep U-Net convolutional networks.in proc. of the International Society for Music Information Retrieval Conference (ISMIR).

[Mohla, 2020] Mohla, S. et al. (2020) FusAtNet: Dual Attention Based SpectroSpatial Multimodal Fusion Network for Hyperspectral and LiDAR Classification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

[Nika, 2021] Jérôme Nika, Jean Bresson. Composing Structured Music Generation Processes with Creative Agents. *2nd Joint Conference on AI Music Creativity (AIMC)*, 2021, Graz, Austria

[Nika, 2017a] Jérôme Nika, Marc Chemillier, Gérard Assayag. ImproteK: introducing scenarios into human–computer music improvisation. *ACM Computers in Entertainment*, 2017.

[Nika, 2017b] Jérôme Nika, Ken Déguernel, Axel Chemla—Romeu–Santos, Emmanuel Vincent, Gérard Assayag. DYCI2 agents: merging the ”free", "reactive", and "scenario–based” music generation paradigms. *International Computer Music Conference*, Oct 2017, Shangai, China.

[Obin, 2011] N. Obin, MeLos: Analysis and Modelling of Speech Prosody and Speaking Style, PhD. Thesis, Ircam-Upmc, 2011.

[Obin, 2018] C. Robinson, N. Obin, A. Roebel. Sequence-to-sequence modelling of F0 for speech emotion conversion, in IEEE International Conference on Audio, Signal, and Speech Processing (ICASSP), 2018.

[Obin, 2021] C. Le Moine, N. Obin and A. Roebel, *Towards end-to-end F0 voice conversion based on Dual-GAN with convolutional wavelet kernels*, 2021 29th European Signal Processing Conference (EUSIPCO), 2021

[Obin, 2022a] Laurent Benaroya, Nicolas Obin, Axel Roebel (2022). Beyond Voice Identity Conversion: Manipulating Voice Attributes by Adversarial Learning of Structured Disentangled Representations.

[Obin, 2022b] Frederik Bous, Laurent Benaroya, Nicolas Obin, Axel Roebel (2022). Voice Reenactment with F0 and timing constraints and adversarial learning of conversions. Soumis à European Signal Processing Conference (EUSIPCO)

[Rachel, 2017] Rachel M. Bittner, Brian McFee, Justin Salamon, Juan Pablo Bello (2017). Deep Salience Representations for f0 Estimation in Polyphonic Music, in proc. of the International Society for Music Information Retrieval Conference (ISMIR).

[Vaswani, 2017] Vaswani, A. et al. (2017) Attention is All you Need, in International Conference on Neural Information Processing Systems (NIPS).