Étude comparative de qualité des codecs audio neuronaux – cas de la musique et du contenu mixte parole/musique

Thomas MULLER[1, 2] Stéphane RAGOT[1] Laetitia GROS[1] Pascal SCALART[2]

[1]Orange Innovation, 2 Avenue Pierre Marzin, 22300 Lannion, France

[2]IRISA, Université de Rennes, 6 Rue de Kerampont, 22300 Lannion, France

Résumé – Cet article présente les résultats d'une caractérisation de la qualité audio des codecs audio neuronaux les plus récents. Les codecs sont évalués sur la musique et le contenu mixte à l'aide d'une méthodologie de test subjectif normalisée. Ces résultats donnent un aperçu de la qualité audio de ces modèles utilisés dans un grand nombre d'applications. La corrélation avec les scores prédits par les modèles objectifs démontre le manque d'outils fiables adaptés à cette nouvelle génération de codecs neuronaux.

Abstract – This article presents the results of an audio quality characterization of state-of-the-art neural audio codecs. The codecs are evaluated on music and mixed content with a standardized subjective test methodology. These results provide an overview of the audio quality of these models, which are now being used in an increasing number of audio applications. The correlation with the quality scores predicted by objective models demonstrates the lack of reliable tools suitable for this new generation of neural codecs.

1 Introduction

L'utilisation des réseaux de neurones artificiels a bouleversé le domaine de la compression audio. De nombreux codeurs-décodeurs (codecs) et vocodeurs "neuronaux" ont été proposés, permettant d'atteindre des débits bien plus faibles qu'en codage "traditionnel". Cette émergence de méthodes neuronales est très souvent motivée par les besoins d'applications comme la synthèse de parole [1], les systèmes de dialogue [2], la traduction de parole [3] ou la génération de musique [4] – un modèle acoustique ou un modèle de langage permet de générer des vecteurs continus ou une représentation discrète (sous forme de "tokens") et un décodeur neuronal sert à générer des signaux audio à partir de cette représentation intermédiaire. Les codecs audio neuronaux actuels, tels que SoundStream [5], EnCodec [6] ou Descript Audio Codec (DAC) [7], s'appuient surtout sur une architecture d'autoencodeur dont l'espace latent est quantifié (comme VQ-VAE [8]) et sont entraînés de façon antagoniste avec des discriminateurs (GAN) pour obtenir de bonnes performances de synthèse audio. Des alternatives comme les modèles de diffusion émergent aussi [9].

La caractérisation des performances de cette nouvelle génération de codecs est essentielle pour comprendre à quel point l'état de l'art est modifié et pour situer ces techniques par rapport aux codecs "traditionnels" encore largement utilisés. Les tests dits subjectifs (ou perceptifs), impliquant des sujets humains évaluant des échantillons sonores selon un protocole prédéfini, restent incontournables pour évaluer la qualité réelle en compression audio. Une caractérisation de la qualité de codecs audio neuronaux a récemment été effectuée pour la parole [10], il n'existe encore pas à notre connaissance d'étude similaire sur la musique et le contenu mixte (mélange parole/musique). Très peu de caractérisations sont proposées dans la littérature, elles se basent généralement sur des prédictions de qualité objectives ou des critères applicatifs [11, 12, 13]. Par ailleurs, de nombreuses métriques objectives existent pour prédire automatiquement la qualité de parole après compression, le choix est plus restreint pour la musique.

L'objectif de cet article est de compléter les études de [10, 14], avec une évaluation de qualité sur la musique et le contenu mixte. Des codecs "traditionnels" sont inclus pour se comparer à l'état de l'art. La corrélation avec les scores prédits par des métriques objectives est analysée pour évaluer la pertinence de ces outils pour cette nouvelle génération de codecs neuronaux.

Les contributions principales de cet article sont :

- La réalisation et la présentation d'un test subjectif pour évaluer les codecs audio neuronaux sur la musique et le contenu mixte (dans le cas monophonique ou mono) ;
- L'étude de la corrélation des scores du test subjectif avec les scores de métriques objectives pour mettre en lumière les limites des outils d'évaluation automatique actuellement disponibles.

2 Codecs testés

La caractérisation a été réalisée avec des codecs disposant d'une implémentation publique. Les codecs choisis sont des codecs optimisés et/ou entraînés pour la musique, fonctionnant à débit fixe (avec un ou plusieurs débits possibles). Une dizaine de codecs neuronaux entraînés uniquement sur de la parole ont également été pré-évalués (par écoute informelle et évaluation objective), mais la dépendance à la base de données d'entraînement était trop forte et la qualité audio trop faible pour être inclus dans le test. La liste complète des codecs et des débits testés est présentée au tableau 1.

2.1 Codecs neuronaux

La majorité des codecs neuronaux sélectionnés sont basés sur une architecture d'autoencodeur convolutif dont l'espace latent est quantifié (VQ-VAE). La quantification est réalisée par quantification vectorielle résiduelle [5] correspondant à la mise en cascade de plusieurs quantificateurs vectoriels. Ce choix permet de modifier le débit du codec en choisissant le nombre de quantificateurs utilisés, permettant d'avoir un seul

TABLEAU 1 : Liste des codecs (mono) et débits testés.

Codec	f_s (kHz)	L (ms)	débit (kbps)	Version
EnCodec	24	13,3	12/24	Nov. 2023
DAC	44,1	11,6	4,3/6/7,8	v1.0.0
HILCodec	24	13,3	4,5/6/9	Oct. 2024
SNAC	44,1	11,6	2,6	v1.2.1
FlowDec	48	13,3	4,5/6/7,5	v0.1 (75m)
xHE-AAC	48	16 à 85,3	8/12/16/24	v4.4.0 FhG Enc.
Opus - audio	48	20	16/24	v1.5.2 (–cbr)
Opus - voip	48	20	12/16/24	v1.5.2 (–cbr)
EVS	32	20	9,6/13,2/24,4	v16.3.0

modèle neuronal dont il est possible de régler le compromis débit/qualité. Ces codecs sont EnCodec [6], Descript Audio Codec (DAC) [7], HILCodec [15] et SNAC [16]. EnCodec est le premier codec de ce type proposant une implémentation open source, avec la particularité d'avoir des couches récurrentes LSTM. DAC est un modèle non causal avec plus de paramètres entraînables qui a également proposé des améliorations pour la quantification de l'espace latent et l'utilisation de la fonction d'activation Snake. HILCodec propose une architecture plus légère de 10M de paramètres (contre 75M pour DAC) et contrôle la variance des activations en ajoutant des normalisations soigneusement choisies. Enfin, SNAC reprend DAC et propose une quantification vectorielle résiduelle multi-échelles censée mieux s'adapter à la nature temporelle multi-échelles des signaux audio.

De plus, nous incluons le codec FlowDec [9] basé sur l'architecture du codec DAC mais utilisant un modèle de diffusion comme post-filtre pour remplacer l'entraînement GAN réputé instable et difficile à maîtriser.

2.2 Codecs traditionnels

Trois codecs traditionnels sont testés : Opus [17], EVS [18] et xHE-AAC. Opus est le standard de l'IETF utilisé dans WebRTC et diverses applications d'appels sur Internet. Il est testé selon deux modes : le mode voix sur IP (VoIP) combinant un codage (SILK) par prédiction linéaire et un codage (CELT) par transformée, et le mode audio s'appuyant uniquement sur le codage par transformée. EVS (pour Enhanced Voice Services) est le standard du 3GPP pour les application de téléphonie mobile, combinant une variété de techniques avancées (à base de prédiction linéaire, codage par transformée, extension de bande). xHE-AAC (pour extended High Efficiency - Advanced Audio Coding) est dérivé de la norme MPEG USAC (Unified Speech and Audio Coding) et est utilisé pour la radiodiffusion numérique ou la diffusion audiovisuelle (par exemple Netflix). Ce sont tous les trois des codecs versatiles multi-débits permettant de compresser la parole et la musique.

3 Test subjectif et métriques objectives

Nous évaluons des codecs neuronaux de l'état de l'art en utilisant la méthodologie de test normalisée P.800 DCR [19], dite d'évaluation par catégories de dégradation (Degradation Category Rating). Plusieurs méthodologies de test subjectif sont possibles. Comme expliqué dans [10], nous utilisons la méthodologie P.800 DCR car elle a fait ses preuves en normalisation 3GPP et permet de comparer de nombreuses conditions (ici 30 conditions) et sur plusieurs bandes audio, avec des sujets "naïfs" (non experimentés ni experts) reflétant la population générale. Il s'agit d'un test de dégradation où les auditeurs naïfs doivent noter la dégradation d'un échantillon audio vis à vis de l'échantillon audio non codé sur une échelle à 5 points – cette échelle est définie sur l'axe des ordonnées à la figure 1.

TABLEAU 2 : Base de test (5+1 échantillons par catégorie).

Catégorie	Contenu	Description
1	musique classique	instrumental : orchestre, piano, clavecin, musique médiévale, métallophone
2		vocal : opéra, chorale, voix à capella
3	musique moderne	instrumental : groupe rock, trompette jazz, harmonica, guitare électrique, castagnettes et guitare acoustique, big band
4		vocal : extraits (Jacques Brel, Tracy Chapman, Sarah McLachlan, etc.)
5	contenu mixte	enregistrement naturel : radio, extrait de film, musique d'attente, commentaires sportifs, publicités
6		mélange artificiel parole + musique (rapport signal à bruit de 10 à 25 dB)

Le test réalisé est composé de 30 conditions : la référence (non-codée), la référence dont la bande audio a été réduite à 12kHz et 16kHz, trois ancres bruitées appelées MNRU (Modulated Noise Reference Units) à différents rapports signal sur bruit, et les 24 combinaisons de codec/débit présentées au tableau 1.

La base audio de test comporte 6 catégories, reportées au tableau 2. Chaque catégorie comporte 5+1 échantillons (+1 pour la phase de familiarisation), chacun d'une dizaine de secondes, échantillonnés à 48kHz. Chaque échantillon est prétraité : filtrage passe-bande 20-20000Hz et normalisation du niveau sonore à -26dB LKFS.

Pour réaliser le test, 30 sujets ont été recrutés, séparés en 5 groupes de 6 auditeurs. Après une phase de familiarisation, chaque auditeur écoute les 30 conditions sur une sous-partie des 30 échantillons de la base de test pour une durée de test d'environ 2h avec des pauses. Au total, chaque condition est écoutée et notée 180 fois, permettant une estimation robuste de la qualité audio.

Les tests subjectifs sont onéreux et longs à mettre en place ; des outils de mesure automatique de la qualité audio, appelés métriques objectives, sont souvent utilisés comme alternative. Pour la musique et le contenu mixte, il existe principalement trois métriques pour prédire la qualité audio normalement évaluée par un test subjectif : PEAQ [20], PEMO-Q [21] et ViSQOL Audio [22]. De plus, par curiosité, la métrique objective POLQA [23] donnant les meilleurs résultats sur la parole [14]

TABLEAU 3 : Métriques objectives testées.

Métrique	Contenu	f_s (kHz)	Version
PEAQ	Audio	48	Basic, AFsp v9r0
PEMO-Q	Audio	48	v1.4.1
ViSQOL-A	Audio	48	v3.3.3
POLQA	Parole	48	v3.0 (MultiDSLA)

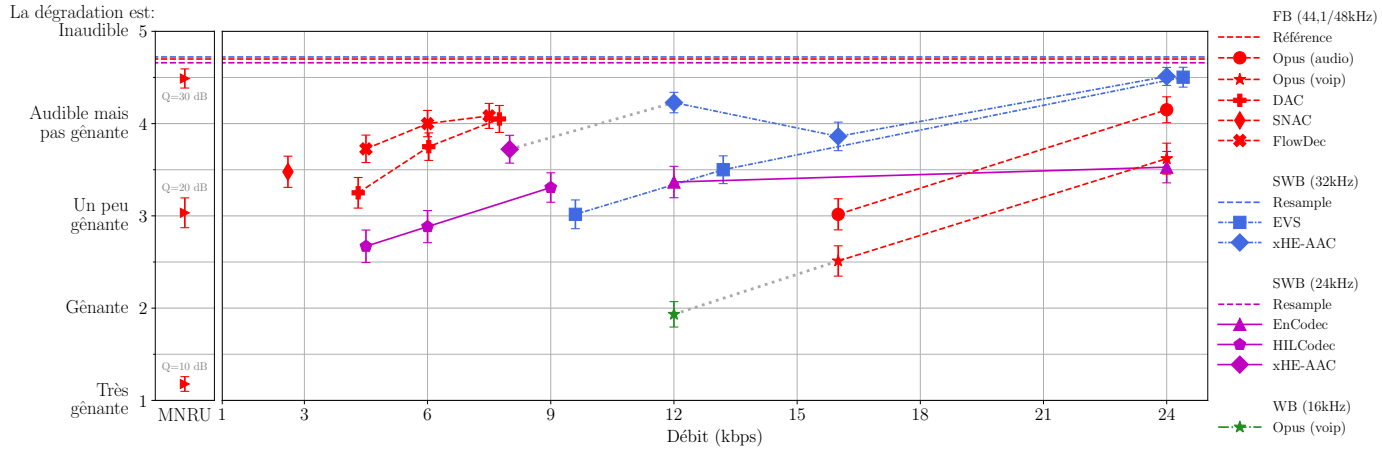

FIGURE 1 : Scores DMOS (Degradation MOS) du test P.800 DCR par condition testée (intervales de confiance à 95%). Les scores de la condition de référence et des conditions à bande audio limitées sont représentés par des lignes horizontales pointillées. Les couleurs indiquent la bande audio de fonctionnement des codecs comparés.

est également testée. Bien qu'elle soit normalement réservée pour la prédiction de qualité pour des signaux de parole, elle a démontré de bonnes prédictions sur les codecs neuronaux [14], et elle s'applique à des signaux en pleine bande (48kHz d'échantillonnage). Elle a également déjà été utilisée hors de son domaine applicatif pour la caractérisation d'une version du codec EVS au 3GPP. Ces quatre métriques objectives sont appliquées à toute la base de test pour comparer les prédictions de qualité aux scores obtenus avec le test subjectif DCR.

4 Résultats et discussion

4.1 Résultats du test subjectif DCR

La figure 1 montre les résultats du test subjectif DCR. Les notes de dégradation DMOS (Degradation Mean Opinion Score) sont tracées en fonction des débits des codecs testés. La position des trois ancres MNRU montre que le test est correctement calibré, et que les sujets ont utilisé toute la plage de notation de 1 à 5. Les codecs atteignant la meilleure qualité audio sont EVS et xHE-AAC pour un débit autour de 24 kbits par seconde (kbps). À ce débit la qualité audio de ces codecs est proche de la saturation et donc de la référence non codée.

Le codec xHE-AAC présente un comportement étonnant, avec une courbe débit-qualité qui n'est pas strictement croissante. Le point à 16 kbps a une qualité bien moindre que celui à 12 kbps. Cela est à priori dû à un mode de fonctionnement différent interne à ce codec à 16 kbps.

L'avantage des codecs neuronaux est clairement visible lorsqu'on observe les plus bas débits. Pour les plus bas débits testés pour EVS et xHE-AAC, les codecs neuronaux présentent de meilleurs scores moyens.

Aucun codec neuronal ne semble arriver proche de la qualité de la référence non codée, avec un plafond à 4,08 DMOS pour FlowDec. Bien que FlowDec affiche les meilleures performances pour les codecs neuronaux, son utilisation reste cependant limitée par la très forte complexité calculatoire (loin du temps-réel) due à l'usage d'un modèle de diffusion.

HILCodec, bien qu'en dessous de FlowDec et DAC, est un codec plus léger que ces deux derniers, comportant bien moins de paramètres entraînables ($\sim$10M contre $\sim$75M pour DAC). Il s'agit donc tout de même d'une alternative intéressante.

Le codec SNAC permet d'atteindre une qualité raisonnable pour un débit de moins de 3kbps, ce qui n'était pas envisageable avec les codecs traditionnels.

4.2 Corrélation avec les scores objectifs

Une étude de corrélation entre les scores prédits par les métriques objectives et les notes du test DCR a été réalisée. Pour comparer les métriques objectives entre elles, trois métriques de corrélation sont utilisées, comme dans [14]. Le coefficient de corrélation de Pearson indique la linéarité de la relation entre les scores prédits et les notes DMOS. La RMSE (racine carrée de l'erreur quadratique moyenne) mesure les différences entre les scores. Le tau de Kendall est un coefficient de corrélation de rang qui mesure la similarité dans l'ordre des conditions, triées par score. De plus, il est d'usage, comme montré dans [14], d'appliquer une fonction de correction ("mapping") sur les scores objectifs pour les rapprocher des notes DMOS et compenser de potentiels biais. Les fonctions de correction couramment utilisées sont des polynômes d'ordre 1 (fonction affine) ou d'ordre 3, en garantissant la monotonie sur l'intervalle de travail $[1, 5]$.

La figure 2 montre ces trois indicateurs de corrélation pour les quatre métriques objectives testées : PEAQ, PEMO-Q, ViSQOL Audio et POLQA. Pour chaque métrique de corrélation, les métriques objectives sont triées par performance décroissante. Nous rappellons que POLQA ne devrait normalement pas être utilisée sur ce type de contenu audio.

Bien que le modèle POLQA ait été testé par curiosité, il s'agit de la métrique qui corrèle le mieux avec les résultats du test DCR selon les coefficients de Pearson et de Kendall et selon la RMSE après correction polynomiale du troisième degré. ViSQOL Audio, PEAQ et PEMO-Q démontrent des performances moyennes, bien qu'étant normalement adaptés à la musique et le contenu mixte. Globalement, les valeurs des métriques de corrélation de ce test sur la musique et le contenu mixte sont plus faibles que celles du test DCR mené sur la parole [10]. Cela confirme que la tâche de prédiction de qualité audio est plus complexe que pour la qualité de la parole, et que les outils disponibles ne sont pas toujours assez fiables, bien que très utilisés pour comparer des codecs dans les communications scientifiques.

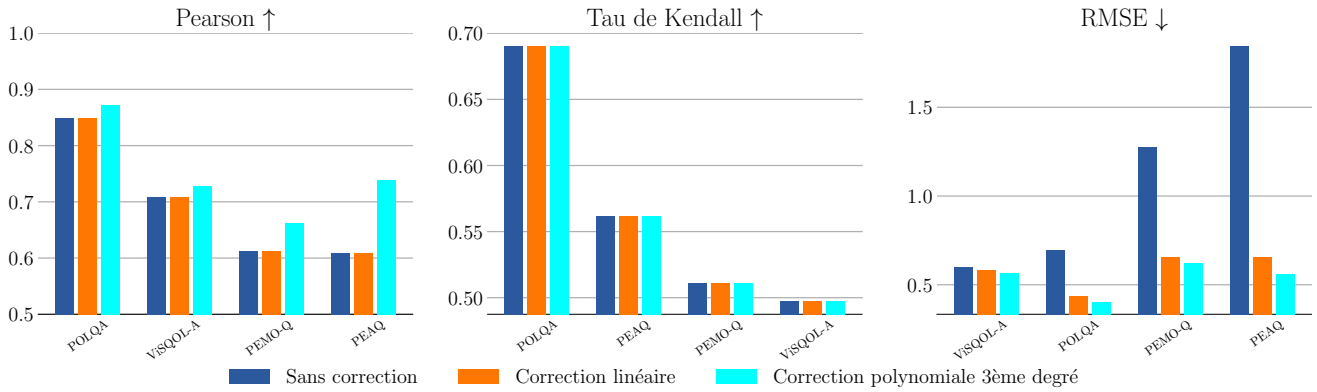

FIGURE 2 : Métriques de corrélation (Pearson, Tau de Kendall et RMSE) entre les notes DMOS du test subjectif DCR et les scores des métriques objectives testées. Les résultats sont montrés sans correction, avec correction linéaire et correction polynomiale du 3ème degré.

5 Conclusion

Une caractérisation des codecs audio neuronaux de l'état de l'art a été réalisée. Cet état des lieux donne un aperçu de la qualité audio atteignable avec cette nouvelle génération de codecs, qui sont aujourd'hui utilisés dans de nombreuses applications audio : synthèse, traduction, génération, etc. Les résultats du test montrent cependant qu'il reste une marge de progression pour obtenir une qualité proche de l'audio de référence (non codée).

De plus, la corrélation des résultats du test subjectif avec les scores prédits par des métriques objectives montre que l'utilisation de tels outils ne reflète pas toujours la qualité perçue, et qu'il manque pour la musique et le contenu audio mixte des métriques permettant une prédiction de qualité fiable. Cela est d'autant plus critique dans les domaines de la génération de musique, où il n'y a pas de référence audio pour se comparer et prédire la qualité de l'audio généré.

Cette étude a quelques limites, en particulier elle ne porte que sur une seule expérience (P.800 DCR). Des tests complémentaires avec des conditions étendues (autres échantillons, pertes de trames, niveaux variés de signal...) pourraient être envisagés dans le futur.

Références

[1] X. TAN : *Neural Text-to-Speech Synthesis*. Springer, 2023.

[2] A. DÉFOSSEZ *et al.* : Moshi : a speech-text foundation model for real-time dialogue. *In arXiv :2410.00037*, 2024.

[3] T. LABIAUSSE *et al.* : High-fidelity simultaneous speech-to-speech translation. *In arXiv :2502.03382*, 2025.

[4] J. COPET *et al.* : Simple and controllable music generation. *In Proc. NeurIPS*, 2023.

[5] N. ZEGHIDOUR *et al.* : SoundStream : An End-to-End Neural Audio Codec. *IEEE/ACM Trans. TASLP*, 2021.

[6] A. DÉFOSSEZ *et al.* : High Fidelity Neural Audio Compression. *Proc. TMLR*, 2023.

[7] R. KUMAR *et al.* : High-fidelity audio compression with improved rvqgan. *In Proc. NeurIPS*, 2023.

[8] A. van den OORD, O. VINYALS et K. KAVUKCUOGLU : Neural discrete representation learning. *In arXiv :1711.00937*, 2018.

[9] S. WELKER *et al.* : Flowdec : A flow-based full-band general audio codec with high perceptual quality. *In Proc. ICLR*, 2025.

[10] T. MULLER *et al.* : Speech quality evaluation of neural audio codecs. *In Interspeech*, 2024.

[11] H. WU *et al.* : Codec-SUPERB : An in-depth analysis of sound codec models. *In Proc. ACL*, 2024.

[12] P. MOUSAVI *et al.* : Dasb–discrete audio and speech benchmark. *In arXiv :2406.14294*, 2024.

[13] J. SHI *et al.* : Espnet-codec : Comprehensive training and evaluation of neural codecs for audio, music, and speech. *In arXiv :2409.15897*, 2024.

[14] T. MULLER *et al.* : Evaluation of objective quality models on neural audio codecs. *In Proc. IWAENC*, 2024.

[15] S. AHN *et al.* : HILCodec : High-Fidelity and Lightweight Neural Audio Codec. *In arXiv :2405.04752*, 2024.

[16] H. SIUZDAK, F. GRÖTSCHLA et L.A. LANZENDÖRFER : SNAC : Multi-Scale Neural Audio Codec. *In arXiv :2410.14411*, 2024.

[17] J.-. VALIN, K. VOS et T.B. TERRIBERRY : Definition of the Opus Audio Codec. RFC 6716, 2012.

[18] M. DIETZ *et al.* : Overview of the EVS codec architecture. *In Proc. ICASSP*, 2015.

[19] ITU-T REC. P.800 : Methods for subjective determination of transmission quality, Aug. 1996.

[20] T. THIEDE *et al.* : PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality. *J. Audio Eng. Soc*, 2000.

[21] R. HUBER et B. KOLLMEIER : PEMO-Q – A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception. *Proc. TASLP*, 2006.

[22] M. CHINEN *et al.* : ViSQOL v3 : An Open Source Production Ready Objective Speech and Audio Metric. *In Proc. QoMEX*, 2020.

[23] ITU-T REC. P.863 : Perceptual objective listening quality prediction, Mar. 2018.