Amélioration de codecs de parole par modèle de diffusion

Romain Buguet* Stéphane Ragot Thomas Muller
Orange Innovation, Lannion
{prenom.nom}@orange.com

Résumé

Dans cet article on étend une méthode de post-traitement appelée SPF (Score-based Post-Filter), utilisant un modèle de diffusion et visant à réduire le bruit de codage audio. On s'intéresse d'abord au post-traitement du codec Opus (à 24 kbit/s) en modifiant le signal d'entrée de référence lors de l'entraînement. Le post-traitement est ensuite étendu au codec AMR-WB dans le cas multi-débits (6,6, 8,85 et 12,65 kbit/s), avec une amélioration de qualité significative.

Mots clefs

Codage audio, amélioration de la parole, modèles de diffusion, évaluation de la qualité audio.

1 Introduction

Les approches traditionnelles de codage audio mono – dont l'état de l'art est représenté par des codecs comme EVS [1] ou Opus [2] pour les applications conversationnelles – sont basées sur une représentation du signal combinant des blocs de traitement de manière experte. Depuis 2017, on observe une émergence de méthodes de bout en bout de codage audio par réseaux de neurones, comme le vocodage basé WaveNet [3], et plus récemment des codecs de type autoencodeurs basés GAN (Generative Adversarial Networks), comme SoundStream [4], AudioDec [5] ou DAC [6], dont certains offrent un compromis débit/qualité jusque-là inatteignable. Les modèles de diffusion commencent à donner des résultats très prometteurs en codage audio [7, 8, 9] ; dans ce type de modèles, la génération consiste à inverser un processus de diffusion stochastique à partir d'un réseau de neurones.

L'objet de cet article est d'étudier et d'étendre la méthode ScoreDec récemment proposée dans [9] pour améliorer la qualité de codecs de parole. Celle-ci consiste à appliquer à la sortie d'un codec de parole existant (AudioDec ou Opus dans [9]) un post-traitement appelé SPF (Score-based Post-Filter) utilisant un modèle de diffusion. Ce post-traitement réutilise en fait la méthode SGMSE (*Score-based Generative Model for Speech Enhancement*) de [10].

Le post-traitement est une méthode classique d'amélioration de qualité après codage avec des approches "traditionnelles" par traitement du signal [11, 12, 13, 14] ou plus récemment par réseaux de neurones [15, 16, 17, 18, 19]. La spécificité de l'approche ScoreDec est d'appliquer un débruitage par modèle de diffusion.

Cet article est organisé comme suit. La méthode ScoreDec est revue à la section 2. L'amélioration de ScoreDec pour Opus et l'extension à AMR-WB sont détaillées à la section 3, avant de présenter les résultats expérimentaux à la section 4 et conclure à la section 5.

*Romain Buguet était en stage de fin d'études quand ce travail a été réalisé.

2 Revue de la méthode ScoreDec

2.1 Modèles de diffusion

Modèles de diffusion Les modèles de diffusion sont une classe de modèles génératifs inspirés de la physique statistique [20] et développés récemment dans le contexte de la génération d'images [20, 21]. Le mécanisme des modèles de diffusion repose sur deux processus stochastiques, l'un vers l'avant, dit *forward*, et l'autre rétrograde, appelé processus *backward*. Au cours du premier, la structure des données, de distribution p inconnue, est progressivement détruite et évolue lentement vers une distribution cible connue et simple à échantillonner, le plus souvent gaussienne. Cette transformation est modélisée par une chaîne de Markov $\{\mathbf{x}_t\}_{t=0}^{T}$ issue de $\mathbf{x}_0 \sim p$, indexée par la variable de temps t, et dont les états correspondent à des niveaux de bruits de plus en plus élevés, si bien que la distribution de l'état final $\mathbf{x}_T$ est proche d'une loi normale centrée réduite. À l'inverse, le processus *backward* génère un échantillon $\mathbf{x}_0 \sim p_\theta$ (où p_θ est proche de p) par débruitages successifs, à partir d'un bruit gaussien $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$. L'apprentissage consiste à estimer le bruit ajouté à chaque étape du processus *forward* afin de le soustraire graduellement lors de l'inférence.

Modèles génératifs basés sur le score Plutôt que d'estimer directement la distribution p des données au moyen d'un modèle p_θ, les modèles génératifs basés sur le score estiment la *fonction de score (de Stein)* $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ des données à l'aide d'un *modèle de score* s_θ [22]. Dans [23], Song et al. proposent le formalisme suivant pour le processus *forward*, basé sur l'équation différentielle stochastique (EDS)

$$\mathrm{d}\mathbf{x}_t = \mathbf{f}(t, \mathbf{x}_t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}_t, \tag{1}$$

où le champ de vecteurs $\mathbf{f}$ est un terme de dérive (*drift*), gouvernant le comportement moyen de l'équation, g est un coefficient contrôlant la quantité de bruit injectée à chaque instant, et $\mathbf{w}_t$ est un mouvement brownien (processus de Wiener). La variable temporelle t évolue ici continûment entre l'instant initial $t = 0$ et l'instant final $t = T$. D'après [24], le processus *backward* correspond à la solution de l'EDS rétrograde associée à (1)

$$\mathrm{d}\mathbf{x}_t = \left[-\mathbf{f}(t, \mathbf{x}_t) + g(t)^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}_t, \tag{2}$$

qui fait intervenir la fonction de score de la distribution p_t des données au temps t, ainsi qu'un processus de Wiener rétrograde $\bar{\mathbf{w}}_t$. Le score est approximé par $s_\theta(\mathbf{x}_t, t)$, modélisé par un réseau de neurones. Une fois le modèle entraîné, de nouveaux échantillons sont produits par simulation du processus *backward* (2) à l'aide de méthodes de résolution numérique d'EDS, après avoir substitué le modèle s_θ à la fonction de score.

2.2 Modèle SGMSE

SGMSE [10] est un modèle génératif basé sur le score pour l'amélioration de la parole. Le signal de parole observé est noté $\mathbf{y}$ et correspond à une version bruitée du signal de parole pure $\mathbf{x}$. Le modèle SGMSE inclut un processus *forward* dont le but est de

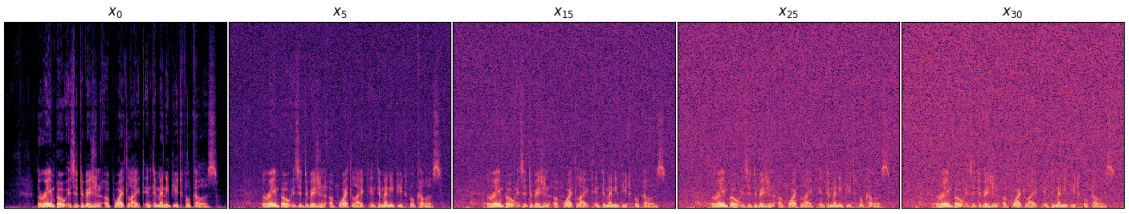

FIGURE 1 – *Spectrogramme de $\mathbf{x}_t$ (ici montré uniquement en amplitude) à différentes itérations du processus* forward.

transférer la distribution - inconnue - de la parole pure $\mathbf{x}$ vers une distribution gaussienne centrée sur les données bruitées (avec une dérive progressive de $\mathbf{x}$ vers $\mathbf{y}$ lors du processus de diffusion), et un processus *backward* réalisant la transformation inverse. Dans ce modèle, les données sont traitées sous forme de spectrogrammes complexes. Avec les notations du paragraphe précédent, le terme de *drift* $\mathbf{f}(t, \mathbf{x}_t)$ de l'équation 1 est remplacé par

$$\mathbf{f}(\mathbf{x}_t, \mathbf{y}) := \gamma(\mathbf{y} - \mathbf{x}_t), \tag{3}$$

et le coefficient de diffusion est donné par

$$g(t) := \sigma_{\min} \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^t \sqrt{2 \log \left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)}, \tag{4}$$

où $\gamma > 0$ est un coefficient de raideur et les hyperparamètres $\sigma_{\min}$ et $\sigma_{\max}$ contrôlent la diffusion. Notons que la fonction $\mathbf{f}$ à l'équation 3 dépend de $\mathbf{y}$ et ne dépend pas explicitement du temps t – pour simplifier les notations, on garde cependant le même symbole $\mathbf{f}(.)$.

Inférence SGMSE. À l'inférence, les audio dégradés $y(n)$ sont normalisés par $M = \max_n |y(n)|$ puis transformés en spectrogrammes complexes $\mathbf{y} \in \mathbb{C}^{K \times F}$ par transformée de Fourier discrète à court-terme (STFT) et compression d'amplitude, où K et F sont les nombres de trames temporelles et de raies fréquentielles. Le modèle de diffusion SGMSE, basé sur l'architecture de réseau de neurones NCSN++ (pour *Noise Conditional Score Network*), synthétise une version améliorée $\hat{\mathbf{x}}$ de $\mathbf{y}$, qui est ensuite convertie dans le domaine temporel et renormalisée. Pour cela, une version fortement corrompue $\mathbf{x}_T$ de $\mathbf{y}$ est échantillonnée selon une distribution gaussienne complexe centrée en $\mathbf{y}$. L'intervalle $[0, T]$ est discrétisé en N sous-intervalles de longueur ΔT, et l'équation *backward* est résolue entre $t = T$ et $t = t_\epsilon$, où $t_\epsilon \simeq 0$, à l'aide de méthodes numériques telles que la méthode d'Euler-Maruyama. La limitation à t_ϵ permet d'éviter des instabilités numériques pouvant se produire pour t proche de 0.

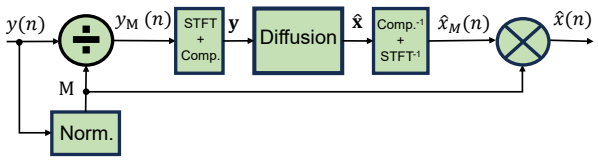

FIGURE 2 – *Modèle SGMSE (inférence).*

Entraînement SGMSE. À chaque étape de l'entraînement, un temps t est d'abord échantillonné selon une distribution uniforme sur l'intervalle $[t_\epsilon, T]$, puis un couple $(\mathbf{x}_0, \mathbf{y})$ de spectrogrammes purs/dégradés est choisi aléatoirement dans la base de données. Connaissant $\mathbf{x}_0$ et $\mathbf{y}$, la distribution de $\mathbf{x}_t$ peut être déterminée explicitement : on peut donc échantillonner directement $\mathbf{x}_t$ et calculer la fonction de score correspondante. Une distance l_2 entre le modèle de score et le score est enfin calculée, puis les paramètres du réseau sont actualisés.

2.3 Post-traitement SPF dans ScoreDec

La méthode ScoreDec [9] est résumée à la figure 3, où un codeur existant comme Opus (incluant un préfiltre passe-haut noté H) prend en entrée le signal $x(n)$, puis le décodeur reconstruit un signal $y(n)$. La méthode ScoreDec revient à post-traiter $y(n)$ avec le modèle de débruitage SGMSE décrit à la figure 2 ; le bruit de codage est ainsi réduit pour améliorer la qualité audio.

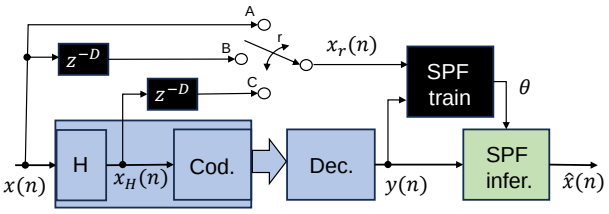

FIGURE 3 – *Modèle ScoreDec avec 3 points de référence ($r = A, B, C$) – $r = A$ correspond au cas traité dans [9].*

Dans [9], le codec Opus est traité comme une "boîte noire" : l'entraînement du post-traitement SPF – qui donne les paramètres θ du modèle de diffusion – est ainsi réalisé en prenant comme point de référence $r = \text{A}$, soit $x_r(n) = x(n)$.

3 Méthode proposée

3.1 Amélioration de SPF pour Opus

Nous proposons d'améliorer ScoreDec, en considérant Opus non pas comme une "boîte noire" mais en prenant en compte des caractéristiques connues du codec. En particulier, Opus induit un retard algorithmique de D échantillons entre l'entrée $x(n)$ et la sortie $y(n)$ – par défaut, à 48 kHz, $D = 312$ (6,5 ms). De plus, Opus inclut un préfiltre passe-haut H (de fréquence de coupure adaptative autour de 60–70 Hz).
On propose donc de modifier la procédure d'entraînement de ScoreDec, comme indiqué à la figure 3 : le signal de référence est remplacé soit par le signal retardé $x_r(n) = x(n - D)$ (cas $r =$ B), soit par le signal préfiltré et retardé $x_r(n) = x_H(n - D)$ (cas $r =$ C) – le signal $y(n)$ en sortie du codec Opus reste le même peu importe le signal d'entrée de référence. Ainsi, le post-filtre SPF ne compense pas le retard algorithmique d'Opus (cas $r =$ B ou C), et ne modélise pas le préfiltre H (cas $r =$ C). La capacité du modèle de diffusion est ainsi concentrée sur la réduction du bruit de codage induit par Opus.

3.2 Extension de ScoreDec à AMR-WB

Nous proposons aussi d'étendre la méthode ScoreDec au codec AMR-WB [25]. Le post-traitement SPF pour AMR-WB suit le

principe de la figure 3, mis à part que les blocs de codage et décodage correspondent respectivement au codeur et décodeur AMR-WB. À la différence d'Opus, on traite le cas multi-débits en considérant les trois modes d'AMR-WB à plus bas débit ($6,6$, $8,85$ et $12,65$ kbits/s) qui sont utilisés en téléphonie mobile. Il est alors possible d'étudier si un post-traitement SPF entrainé à un débit donné reste optimal pour un autre débit d'AMR-WB.
À noter que le retard algorithmique d'AMR-WB est de $D = 95$ échantillons à 16 kHz. AMR-WB inclut aussi un préfiltre passe-haut (avec une fréquence de coupure de 31 Hz). Cependant, ce préfiltre opère sur une bande basse 0–6,4 kHz après décimation à 12,8 kHz ; pour simplifier, on se restreint au cas $r = \text{B}$ de la figure 3 pour AMR-WB.

4 Expériences

4.1 Protocole expérimental

Bases de données : Pour Opus, le modèle ScoreDec est entraîné et testé sur la même base de données Valentini [26] à 48 kHz (parole pure uniquement) que dans [9], pour des questions de reproductibilité et de comparaison directe avec [9]. Pour AMR-WB, cette base Valentini est ré-échantillonnée à 16 kHz.

Paramètres des modèles testés : La méthode ScoreDec pour Opus est configurée par défaut comme dans [9] (avec des trames de 320 échantillons). Pour AMR-WB, on reprend les paramètres SGMSE de [10] (avec des trames de 128 échantillons).

Evaluation de qualité : Les intervalles de confiance à 95% sont donnés avec chaque note moyenne de qualité. La qualité est évaluée par les métriques objectives PESQ [27] et SI-SDR [28] comme dans [9]. Le SI-SDR s'interprète comme un rapport signal à bruit, mais il est difficile à corréler avec une évaluation subjective. Le score PESQ prédit la note MOS (Mean Opinion score) sur l'échelle de qualité d'écoute : $1\rightarrow$ mauvaise, $2\rightarrow$ médiocre, $3\rightarrow$ passable, $4\rightarrow$ bonne, $5\rightarrow$ excellente.

Configuration matérielle : Les entraînements et inférences de SPF sont réalisés sur un GPU A100 avec 40 Go de RAM. Un entraînement SPF dure en moyenne 96 h pour Opus et 30 h pour AMR-WB (sur la base Valentini). L'inférence sur la base de test Valentini prend 3 h pour Opus et 45 min pour AMR-WB (pour 45 min de parole).

4.2 Résultats pour Opus

Les résultats de qualité objective pour Opus (à 24 kbit/s) sont résumés dans le tableau 1. Le cas $r = \text{A}$ du post-traitement SPF (noté Opus_SPF) correspond à [9] – les poids pré-entrainés du modèle de [9] n'étant pas disponibles, ce modèle a été ré-entrainé mais nos résultats sont très proches de ceux de [9] dans les mêmes conditions. Les résultats de [9] sont rappelés au tableau 1 à des fins de comparaison. Notre version ré-entrainée d'Opus_SPF donne des résultats objectifs légèrement meilleurs que ceux présentés dans [9]. La seule différence significative concerne le score SI-SDR pour Opus. Contrairement à [9], la qualité d'Opus est ici mesurée en prenant le signal $x_H(n - D)$ comme signal d'entrée d'Opus avant calcul de la métrique SI-SDR ; le SI-SDR passe ainsi de $-20,62$ dB dans [9] à $12,48$ dB pour la même base de test ; cela élimine le biais dû au retard algorithmique et au préfiltre, on observe un gain de seulement $4,01$ dB en SI-SDR entre Opus et Opus_SPF avec $r = \text{A}$; on confirme une amélioration du score PESQ d'environ 0.09 pour Opus_SPF avec $r = \text{A}$.
Les scores PESQ d'Opus_SPF modifié selon nos propositions à la section 3.1 sont légèrement meilleurs en changeant l'entraînement du modèle (en passant à $r = \text{B}$ puis C) – le SI-SDR montre un gain plus net.

Codec	r	SI-SDR (dB)	PESQ (MOS-LQO$_{wb}$)
Opus [9]	–	-20,62	4,21
Opus (nos tests)	–	$12,48 \pm 0,09$	$4,24 \pm 0,01$
Opus_SPF [9]	A	16.20	4.29
Opus_SPF (nos tests)	A	$16,49 \pm 0,12$	$4,33 \pm 0,02$
	B	$16,66 \pm 0,12$	$4,34 \pm 0,01$
	C	$\mathbf{17,08 \pm 0,13}$	$\mathbf{4,35 \pm 0,02}$

TABLEAU 1 – *Qualité objective sur la base de test Valentini pour Opus à 24 kbit/s avec ou sans post-traitement SPF (avec un point de référence r = A, B ou C à l'entraînement tel que défini à la figure 3).*

Un test d'écoute de type "RefAB" a été réalisé par 5 sujets experts, avec 30 double phrases de parole (8s) en français (normalisées en niveau d'écoute). Ce test subjectif a confirmé les scores PESQ : il n'y pas de différence statistiquement significative entre Opus_SPF avec $r = \text{A}$ et $r = \text{C}$.
Cette étude sur Opus a repris les conditions de [9] où Opus opère à 24 kbit/s, alors qu'Opus a déjà une bonne qualité sur la parole pure à ce débit. Par la suite, il serait intéressant de tester Opus à un débit inférieur à 24 kbit/s.

4.3 Résultats pour AMR-WB

Les résultats de qualité objective pour AMR-WB (à 6,6, 8,85 et 12,65 kbit/s) sont résumés au tableau 2 et illustrés à la figure 4. Le post-traitement SPF pour AMR-WB (abrégé en SPF_x, où x est le débit d'entraînement) est entraîné séparément pour chaque mode (x = 6,6, 8,85 et 12,65), puis sur une base de données comportant un mélange de ces trois débits – ce dernier cas est noté "SPF_MR" (MR pour Multi-Rate).
La comparaison inclut AMR-WB et sa version post-traitée, SPF_x ou SPF_MR. Le post-traitement SPF_x entraîné pour AMR-WB à x kbit/s améliore significativement le score PESQ au même débit (avec un écart de 0,80, 0,52 et 0,37 à 6,6, 8,85 et 12,65 kbit/s, respectivement) ; cette nette amélioration est confirmée par des écoutes informelles sur la base de test Valentini.
Comme on pourrait l'attendre, le post-traitement SPF_x est surtout optimisé pour le débit pour lequel il a été entraîné. En termes de score PESQ, le modèle donnant le meilleur résultat à un débit donné est celui qui a été entrainé pour ce même débit ; c'est également le cas en termes de SI-SDR, sauf au débit de 6,6 kbit/s où le modèle à 8,85 kbit/s obtient un score SI-SDR numériquement supérieur – on note cependant que les modèles entrainés à 6,6, 8,85 et 12,65 kbit/s sont en fait équivalents (en tenant compte de la marge d'erreur). La métrique SI-SDR est connue pour être faiblement corrélée à une évaluation subjective (perceptive), on considère ici PESQ comme donnant une meilleure indication (prédiction) de la qualité perçue du signal post-traité.
Parmi les méthodes précédemment proposées de post-traitement neuronal, on retient ici la méthode par réseaux convolutionnels de [15] qui s'applique aussi à AMR-WB ; dans le tableau IV de [15] (avec une base en anglais et allemand) les tests limités au débit de 12,65 kbit/s donnent un score PESQ de 3,60 pour AMR-WB et 3,85 pour la meilleure variante de post-traitement, soit un gain de 0,25 en termes de score PESQ. Ici, pour le même débit de 12,65 kbit/s le tableau 2 montre un score PESQ de 3,58 pour AMR-WB et 3,95 pour SPF_12,65, soit un gain de 0,37 en termes de score PESQ. Même si ces résultats ne sont pas directement comparables car ils ne sont pas obtenus dans les mêmes conditions

de tests, on peut s'attendre à ce que le débruitage par modèle de diffusion donne de meilleurs résultats étant donné que l'approche considérée de type SGMSE relâche certaines contraintes (étant plus complexe et non causale).

Le post-traitement SPF_MR entraîné en multi-débits a une performance plus homogène que SPF_x mais sous-optimale. Par la suite, il sera intéressant de donner le débit de décodage comme contexte (conditionnement) au modèle de diffusion pour viser une performance optimale en multi-débits. Par ailleurs, on pourra aussi ajouter à la comparaison du tableau 2 plus de débits d'AMR-WB (par exemple le débit maximal de 23,85 kbit/s aussi utilisé en téléphonie) et comparer la qualité avec le décodage AMR-WB amélioré dans EVS (EVS-IO pour InterOperable) [1] qui utilise des techniques classiques de post-traitement [14].

Codec	Mode	SI-SDR (dB)	PESQ (MOS-LQO$_{wb}$)
AMR-WB		$6,00 \pm 0,10$	$2,60 \pm 0,02$
SPF_6,6		$11,30 \pm 0,09$	$\mathbf{3,40 \pm 0,02}$
SPF_8,85	6,6	$\mathbf{11,38 \pm 0,10}$	$3,35 \pm 0,02$
SPF_12,65		$11,34 \pm 0,09$	$3,29 \pm 0,02$
SPF_MR		$10,95 \pm 0,10$	$3,31 \pm 0,02$
AMR-WB		$7,20 \pm 0,11$	$3,08 \pm 0,03$
SPF_6,6		$12,17 \pm 0,02$	$3,54 \pm 0,02$
SPF_8,85	8,85	$13,64 \pm 0,10$	$\mathbf{3,71 \pm 0,03}$
SPF_12,65		$\mathbf{13,71 \pm 0,10}$	$3,66 \pm 0,02$
SPF_MR		$12,87 \pm 0,10$	$3,61 \pm 0,02$
AMR-WB		$7,99 \pm 0,11$	$3,58 \pm 0,03$
SPF_6,6		$12,50 \pm 0,10$	$3,58 \pm 0,03$
SPF_8,85	12,65	$13,65 \pm 0,11$	$3,73 \pm 0,02$
SPF_12,65		$\mathbf{15,38 \pm 0,11}$	$\mathbf{3,95 \pm 0,02}$
SPF_MR		$14,82 \pm 0,11$	$3,90 \pm 0,02$

TABLEAU 2 – *Qualité objective sur la base de test Valentini pour AMR-WB, SPF_x (où x est le débit d'entraînement) ou SPF_MR (pour Multi-Rate) – le post-traitement SPF est entraîné avec le point de référence $r = B$ tel que défini à la figure 3.*

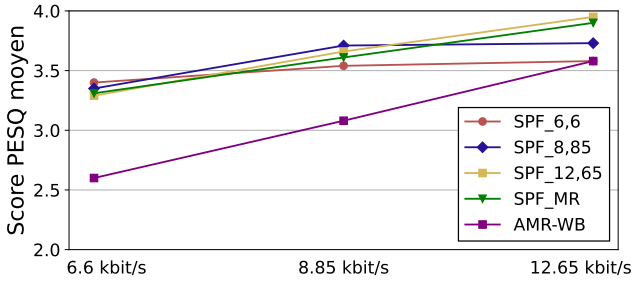

FIGURE 4 – *Représentation graphique des résultats pour AMR-WB donnés au tableau 2.*

5 Conclusion

Dans cet article, la méthode de post-traitement SPF de [9] a été améliorée pour Opus et étendue à AMR-WB. Le gain en qualité apporté par le post-traitement SPF est prometteur et significatif pour AMR-WB, cependant cette méthode traite des séquences audio complètes (et non par trame), avec une complexité très élevée et sans opérer en temps réel par trame courte (≤ 20 ms) ; il sera intéressant de corriger ces aspects en modifiant le modèle SGMSE sous-jacent. L'approche supervisée de [29] pourra aussi être testée dans le contexte du codage. L'extension de cette étude à d'autres contenus audio que la parole (pure) sera aussi nécessaire.

Références

[1] M. Dietz et al. Overview of the EVS codec architecture. Dans *Proc. ICASSP*, 2015.

[2] IETF RFC 6716. Definition of the Opus Audio Codec, 2012.

[3] W.B. Kleijn et al. Wavenet based low rate speech coding. arXiv :1712.01120, 2017.

[4] N. Zeghidour et al. SoundStream : An End-to-End Neural Audio Codec. *IEEE/ACM TASLP*, 30, 2021.

[5] Y.-C. Wu et al. Audiodec : An Open-Source Streaming High-Fidelity Neural Audio Codec. Dans *Proc. ICASSP*, 2023.

[6] R. Kumar et al. High-Fidelity Audio Compression with Improved RVQGAN. Dans *Proc. NIPS*, 2023.

[7] R. San Roman et al. From discrete tokens to high-fidelity audio using multi-band diffusion. arXiv :2308.02560, 2023.

[8] Y. Haici et al. Generative de-quantization for neural speech codec via latent diffusion. Dans *Proc. ICASSP*, 2024.

[9] Y.-C. Wu et al. ScoreDec : A Phase-Preserving High-Fidelity Audio Codec with a Generalized Score-Based Diffusion Post-Filter. Dans *Proc. ICASSP*, 2024.

[10] J. Richter et al. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM TASLP*, 31 :2351–2364, 2023.

[11] V. Ramamoorthy et N. S. Jayant. Enhancement of ADPCM speech by adaptive postfiltering. *AT&T Bell Laboratories Technical Journal*, 63(8) :1465–1475, 1984.

[12] J.H. Chen et A. Gersho. Adaptive postfiltering for quality enhancement of coded speech. *IEEE Transactions on Speech and Audio Processing*, 3(1) :59–71, 1995.

[13] J.-L. Garcia, C. Marro, et B. Kövesi. A PCM coding noise reduction for ITU-T G.711.1. Dans *Proc. Interspeech*, 2008.

[14] T. Vaillancourt, R. Salami, et M. Jelínek. New post-processing techniques for low bit rate CELP codecs. Dans *Proc. ICASSP*, 2015.

[15] Z. Zhao, H. Liu, et T. Fingscheidt. Convolutional Neural Networks to Enhance Coded Speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(4) :663–678, 2019.

[16] Srikanth Korse, Kishan Gupta, et Guillaume Fuchs. Enhancement of Coded Speech Using a Mask-Based Post-Filter. Dans *Proc. ICASSP*, 2020.

[17] Kishan Gupta, Srikanth Korse, Bernd Edler, et Guillaume Fuchs. A DNN Based Post-Filter to Enhance the Quality of Coded Speech in MDCT Domain. Dans *Proc. ICASSP*, 2022.

[18] S. Korse, N. Pia, K. Gupta, et G. Fuchs. PostGAN : A GAN-Based Post-Processor to Enhance the Quality of Coded Speech. Dans *Proc. ICASSP*, 2022.

[19] J. Büthe, J.-M. Valin, et A. Mustafa. Lace : A Light-Weight, Causal Model for Enhancing Coded Speech Through Adaptive Convolutions. Dans *Proc. WASPAA*, 2023.

[20] J. Sohl-Dickstein et al. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. Dans *Proc. ICML*, 2015.

[21] J. Ho et al. Denoising diffusion probabilistic models. Dans *Proc. NeurIPS*, 2020.

[22] Y. Song et al. Generative modeling by estimating gradients of the data distribution. Dans *Proc. NeurIPS*, 2019.

[23] Y. Song et al. Score-based generative modeling through stochastic differential equations. Dans *Proc. ICLR*, 2021.

[24] B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their App.*, 12(3) :313–326, 1982.

[25] B. Bessette et al. The adaptive multirate wideband speech codec (AMR-WB). *IEEE TSAP*, 10(8) :620–636, 2002.

[26] Valentini-Botinhao C. et al. Noisy speech database for training speech enhancement algorithms and TTS models. https ://datashare.ed.ac.uk/handle/10283/2791, 2017.

[27] A.W. Rix et al. Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs. Dans *Proc. ICASSP*, 2001.

[28] J. Le Roux et al. SDR – Half-baked or Well Done ? Dans *Proc. ICASSP*, 2019.

[29] J.E. Ayilo et al. Diffusion-based speech enhancement with a weighted generative-supervised learning loss. Dans *Proc. ICASSP*, 2024.