

Désentrelacement fréquentiel doux pour les codecs audio neuronaux

Benoît GINIÈS Xiaoyu BIE Olivier FERCOQ Gaël RICHARD

LTCI, Télécom Paris, Institut polytechnique de Paris, Palaiseau, France

Résumé – Bien que les modèles basés sur les réseaux de neurones aient permis des avancées significatives dans l’extraction de représentations audio, l’interprétabilité des représentations apprises reste un défi majeur. Pour y remédier, des techniques de désentrelacement ont été intégrées dans les codecs audio neuronaux discrets afin d’imposer une structure aux tokens extraits. Cependant, ces approches sont souvent fortement dépendantes de tâches ou d’ensembles de données spécifiques. Dans ce travail, nous proposons un codec audio neuronal désentrelacé qui tire parti de la décomposition spectrale des signaux temporels pour améliorer l’interprétabilité de la représentation. Des évaluations expérimentales démontrent que notre méthode surpasse un modèle de référence en termes de fidélité de reconstruction et de qualité perceptuelle.

Abstract – While neural-based models have led to significant advancements in audio feature extraction, the interpretability of the learned representations remains a critical challenge. To address this, disentanglement techniques have been integrated into discrete neural audio codecs to impose structure on the extracted tokens. However, these approaches often exhibit strong dependencies on specific datasets or task formulations. In this work, we propose a disentangled neural audio codec that leverages spectral decomposition of time-domain signals to enhance representation interpretability. Experimental evaluations demonstrate that our method surpasses a state-of-the-art baseline in both reconstruction fidelity and perceptual quality.

Ce travail a été financé par l’Union européenne (ERC, HI-Audio, 101052978). Les points de vue et les opinions exprimés sont ceux des auteurs et ne reflètent pas nécessairement ceux de l’Union européenne ou du Conseil européen de la recherche. Ni l’Union européenne ni l’organisme subventionnaire ne peuvent en être tenus pour responsables.

1 Introduction

En traitement audio, l’extraction d’une représentation efficace du signal est essentielle. L’objectif principal d’une telle extraction est d’encoder les informations les plus pertinentes sous une forme compacte. Traditionnellement, ce compromis entre exhaustivité des données et efficacité de la représentation a conduit au développement de représentations conçues à la main pour capturer les propriétés significatives du signal. Les représentations temps-fréquence y jouent un rôle crucial, avec des approches largement reconnues allant de la transformée de Fourier à court terme (TFCT) à des représentations perceptives exploitant l’échelle Mel par exemple. [1]

L’utilisation de réseaux neuronaux pour l’extraction de représentations dans le traitement de l’image et de l’audio a connu un essor important avec l’introduction de l’architecture encodeur-décodeur, appliquée dans le modèle VAE [2], puis étendue avec une étape de quantification dans le modèle VQ-VAE [3]. En prolongement, le quantificateur vectoriel résiduel (RVQ) utilisé dans [4, 5], est venu améliorer la reconstruction. Une relaxation de la quantification a été étudiée et diverses structures de quantificateurs ont été développées dans [6]. L’utilisation de représentations discrètes présente un intérêt notable pour la synthèse vocale [7] ou le transfert de timbre musical [8] par exemple.

Le succès des codecs audio neuronaux a suscité un intérêt croissant pour le désentrelacement des représentations latentes, avec des méthodes souvent adaptées à des applications spécifiques. Par exemple, pour la conversion de la voix chantée, Takahashi & al. [9] intègre le module VQ avec des codeurs de hauteur et d’amplitude. En synthèse vocale, Adam & al. [10]

renforce le désentrelacement du contenu, de la hauteur et de l’identité du locuteur grâce à des structures d’extraction de représentations spécialisées, tandis que Ju & al. [11] allouent des quantificateurs pour capturer la prosodie, le contexte et les détails acoustiques de la parole, complétés par un encodeur de timbre. L’entraînement de ces quantificateurs est guidé par des tâches de supervision appropriées, idée qui a également été appliquée à la séparation des sources [12] par exemple. Bien que nombre de ces approches soient spécifiques à une application, certaines stratégies plus générales ont été proposées. Par exemple, [13] décrit une méthode visant à encourager le désentrelacement des représentations latentes discrètes, et Luo & al. [14] présentent une approche générique avec un codec multibande, où chaque bande de fréquence du spectrogramme d’entrée est traitée séparément.

Dans cet article, nous présentons un nouveau codec neuronal qui génère une représentation discrète et désentrelacée en fréquence. Le modèle de codec proposé fonctionne sur plusieurs fréquences d’échantillonnage, ce qui permet d’extraire une représentation composée de tokens discrets, chacun correspondant à des bandes de fréquence prédéfinies. Nous démontrons que cette représentation désentrelacée apporte une amélioration par rapport au modèle de base et ouvre de nouvelles possibilités pour l’interprétabilité du modèle.

2 Codec désentrelacé

Notre modèle présente une approche générique, facilement applicable, basée sur la décomposition fréquentielle des signaux audio, ce qui lui confère une grande interprétabilité. Notre approche consiste à adapter l’architecture d’un codec audio neuronal discret pour introduire une information fréquentielle dans la représentation extraite. Pour ce faire, nous divisons les spectrogrammes d’entrée en sous-bandes, qui sont ensuite traitées indépendamment (comme dans Luo & al. [14]). Toutefois,

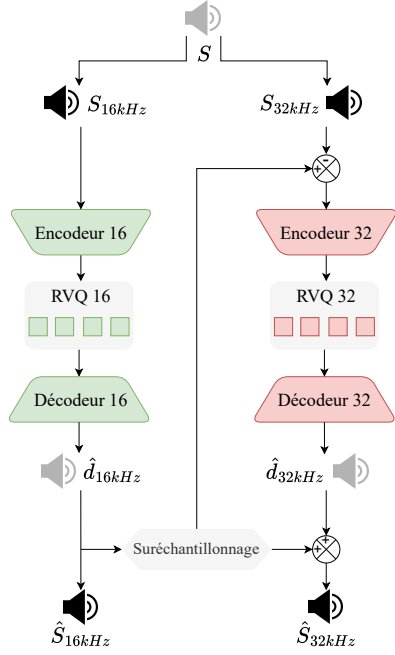


FIGURE 1 : Le codec désentrelacé proposé. La branche 16 kHz reconstruit le signal $[0 - 8\text{ kHz}]$. La branche 32 kHz traite le résidu de $S_{32\text{ kHz}}$ et $U(\hat{d}_{16\text{ kHz}})$ pour produire le signal $[0 - 16\text{ kHz}]$ en additionnant les sorties de chaque branche.

cette méthode peut imposer des contraintes importantes, car chaque bande doit être reconstruite séparément et la combinaison des sous-bandes en un spectrogramme complet peut introduire des artefacts indésirables.

Notre approche n’y est pas sujette puisqu’elle s’appuie sur une décomposition douce des fréquences en travaillant dans le domaine temporel, et en exploitant le lien entre la fréquence d’échantillonnage d’un signal et son information spectrale. Nous ne considérerons qu’une version simple de notre approche avec deux branches (voir la figure 1 pour une description schématique). Cette architecture simplifiée peut servir de preuve de concept et être facilement étendue à plusieurs branches fonctionnant chacune à des fréquences d’échantillonnage différentes. Les deux branches de notre codec audio neuronal se rapportent aux fréquences d’échantillonnages 16 kHz et 32 kHz . Les signaux temporels d’entrée $(S_{16\text{ kHz}}, S_{32\text{ kHz}})$ sont deux versions du même signal S , où $S_{16\text{ kHz}}$ est échantillonné à 16 kHz (c’est-à-dire qu’il contient des informations dans la bande $[0 - 8\text{ kHz}]$) et $S_{32\text{ kHz}}$ est échantillonné à 32 kHz (c’est-à-dire qu’il contient des informations dans la bande $[0 - 16\text{ kHz}]$). Chaque branche produit un signal temporel, noté respectivement $\hat{d}_{16\text{ kHz}}$ et $\hat{d}_{32\text{ kHz}}$. Le signal reconstruit de $S_{16\text{ kHz}}$, noté $\hat{S}_{16\text{ kHz}}$, est directement obtenu à partir de la branche 16 kHz , et : $\hat{S}_{16\text{ kHz}} = \hat{d}_{16\text{ kHz}}$. L’autre branche, qui fonctionne à 32 kHz , prend en entrée le signal résiduel $S_{32\text{ kHz}} - U(\hat{d}_{16\text{ kHz}})$ où U est l’opérateur de suréchantillonnage de 16 kHz à 32 kHz . Le signal reconstruit de cette branche, noté $\hat{S}_{32\text{ kHz}}$, est obtenu en additionnant directement les signaux issus des deux décodeurs : $\hat{S}_{32\text{ kHz}} = U(\hat{d}_{16\text{ kHz}}) + \hat{d}_{32\text{ kHz}}$.

L’architecture en cascade proposée permet un désentrelacement doux : chaque branche est principalement conçue pour

générer du contenu dans sa propre bande de fréquence mais aucune contrainte stricte n’impose cette séparation. Par conséquent, le décodeur de la branche 32 kHz peut également reconstruire du contenu à basse fréquence. Ce désentrelacement doux permet à la branche 32 kHz d’améliorer la reconstruction du signal en dessous de 8 kHz et d’atténuer les artefacts potentiels aux frontières entre les bandes de fréquences. D’autre part, une représentation issue d’une telle architecture peut démontrer un intérêt certain pour une tâche telle que l’extension de bande.

3 Expériences

3.1 Données

Pour nos expériences, nous avons utilisé les bases de données MUSDB18 [15] et Jamendo [16] qui rassemblent plus de 55 000 pistes de musique dans un jeu d’entraînement et 50 pistes musicales dans un jeu de test, initialement échantillonnées à $44,1\text{ kHz}$ avec une durée totale supérieure à 3 700 heures.

3.2 Architecture

Nous avons choisi de reproduire l’architecture du Descript Audio Codec (DAC [5]), un état de l’art en matière de compression audio, dans une version dégradée (c’est-à-dire avec un RVQ ne contenant que quelques dictionnaires, ou en d’autres termes, avec un taux de compression élevé). Il se compose d’un codeur-décodeur convolutionnel, dans lequel est inséré un simple quantificateur vectoriel résiduel. Pour chacune des deux bandes de fréquences définies précédemment, nous avons reproduit une version de ce codec audio et adapté ses dimensions à la fréquence d’échantillonnage correspondante : nous avons choisi les rapports de compression des blocs encodeurs dans chaque branche de manière à ce que le rapport entre la fréquence d’échantillonnage de la branche et le taux de compression global soit constant d’une branche à l’autre. Nous avons décidé de refléter la largeur de chaque bande de fréquence dans la profondeur du RVQ dans chaque branche. Cela a conduit à avoir quatre quantificateurs dans la branche 16 kHz et quatre autres quantificateurs dans la branche 32 kHz .

Pour la comparaison, nous avons également réentraîné le modèle DAC, uniquement sur MUSDB18 et Jamendo, à des fréquences d’échantillonnage de 16 kHz et 32 kHz en conservant les taux de compression et le débit binaire que nous avons définis dans notre modèle (il s’agit donc d’une version dégradée) : un débit binaire de 2 kbps pour le modèle de 16 kHz (taux de compression de 128) et un débit binaire de 4 kbps pour le modèle de 32 kHz (taux de compression de 128).

3.3 Procédure d’entraînement

Le codec ayant plusieurs branches, une légère adaptation de la procédure d’entraînement du modèle DAC [5] a été nécessaire. Toutes les branches sont d’abord entraînées en cascade, en commençant par la branche fonctionnant à la fréquence d’échantillonnage la plus basse : a) nous entraînons une branche, b) nous gelons ensuite ses paramètres et ceux de toutes les branches entraînées précédemment, et enfin c) nous entraînons la branche suivante. À la fin de cette formation en cascade, tous les poids de toutes les branches sont finement

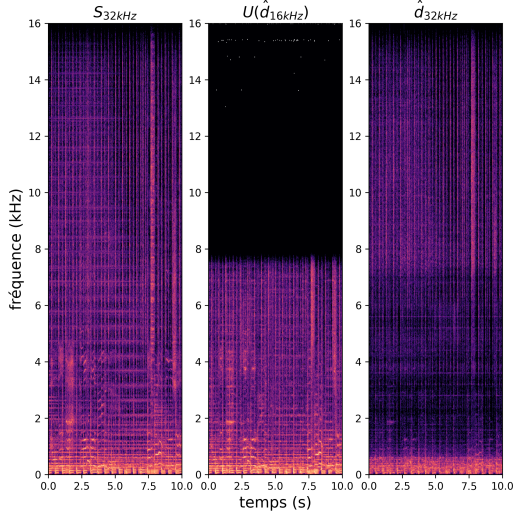


FIGURE 2 : Spectrogrammes de S_{32kHz} , $U(\hat{d}_{16kHz})$ et $\hat{d}_{32kHz}$.

$U(\hat{d}_{16kHz})$ n'encode que l'information dans la bande $[0 - 8 kHz]$. $\hat{d}_{32kHz}$ a la plus grande partie de son énergie dans la bande $[8 - 16 kHz]$, même s'il porte également des informations résiduelles dans la bande inférieure.

ajustés pour trouver un optimum global. Pour chaque branche associée à une fréquence d'échantillonnage F , nous lui adjoignons un réseau discriminant (en conservant la structure du discriminateur décrit dans [5]), et nous calculons les coûts génératifs adverses ($\mathcal{L}_F^{gen}$) et les coûts de *matching features* adverses ($\mathcal{L}_F^{fm}$), un coût de mel multi-échelles ($\mathcal{L}_F^{mel}$) et les coûts des dictionnaires ($\mathcal{L}_F^{cb}$) et un coût d'adhérence ($\mathcal{L}_F^{cmt}$) pour chaque branche.

Pour les premières étapes d'apprentissage 16 kHz et 32 kHz, nous définissons les coûts d'entraînement $\mathcal{L}_{16kHz}^{total}$ et $\mathcal{L}_{32kHz}^{total}$:

$$\mathcal{L}_{16kHz}^{total} = \sum_{\lambda} (\alpha_{16kHz}^{\lambda} \mathcal{L}_{16kHz}^{\lambda}) \quad (1)$$

$$\mathcal{L}_{32kHz}^{total} = \sum_{\lambda} (\alpha_{32kHz}^{\lambda} \mathcal{L}_{32kHz}^{\lambda}) \quad (2)$$

Où $\lambda \in \{gen, fm, mel, cb, cmt\}$. Nous avons choisi de définir le coût d'entraînement pour l'étape d'ajustement final ($\mathcal{L}_{finetun}^{total}$) de la manière suivante :

$$\begin{aligned} \mathcal{L}_{finetun}^{total} = & \sum_{\lambda \in \{gen, fm, mel\}} \frac{1}{2} (\alpha_{32kHz}^{\lambda} \mathcal{L}_{32kHz}^{\lambda} + \alpha_{16kHz}^{\lambda} \mathcal{L}_{16kHz}^{\lambda}) \\ & + \sum_{\lambda \in \{cb, cmt\}} (\alpha_{32kHz}^{\lambda} \mathcal{L}_{32kHz}^{\lambda} + \alpha_{16kHz}^{\lambda} \mathcal{L}_{16kHz}^{\lambda}) \end{aligned} \quad (3)$$

Nous calculons donc la moyenne des coûts génératifs, de feature matching et mel des deux branches, tandis que nous sommions simplement les coûts relatifs à la quantification, pour tenir compte de la différence de nature entre les coûts liés à la reconstruction (pour lesquelles les deux branches se chevauchent) et les coûts qui contraignent l'apprentissage des dictionnaires 16 kHz et 32 kHz.

Nous avons choisi pour $F \in \{16kHz, 32kHz\}$, $\alpha_F^{gen} = 1.0$, $\alpha_F^{fm} = 2.0$, $\alpha_F^{mel} = 15.0$, $\alpha_F^{cb} = 1.0$ et $\alpha_F^{cmt} = 0.25$, comme spécifié dans [5].

TABLE 1 : Métriques de reconstruction du codec désentrelacé

Échantillonnage	16000 Hz		32000 Hz	
Débit	2 kbps		4 kbps	
Modèle	DAC	CD (proposé)	DAC	CD (proposé)
mel ($\downarrow$)	1.08	0.95	0.90	0.80
stft ($\downarrow$)	2.67	2.52	2.28	2.14
reconstruction ($\downarrow$)	0.072	0.066	0.060	0.05
SI-SDR ($\uparrow$)	2.97	3.90	5.00	6.05
ViSQOL ($\uparrow$)	4.08	4.25	3.97	4.22

TABLE 2 : Notes moyennes de qualité perceptive obtenues par le test MUSHRA ($\pm$ écart type)

	Référence	DAC	CD (proposé)	Ancre
16 kHz	95 $\pm$ 9	31 $\pm$ 18	53 $\pm$ 20	37 $\pm$ 25
32 kHz	96 $\pm$ 5	49 $\pm$ 20	66 $\pm$ 19	38 $\pm$ 24

4 Evaluation

Nous avons également conservé les mêmes mesures d'évaluation que dans [5] : coût de forme d'onde, coût TFTC, coût mel, rapport signal-distorsion échelle invariant (SI-SDR) tel qu'introduit dans [17], et le score ViSQOL, une évaluation de la qualité audio perceptive introduite dans [18].

L'informativité de la reconstruction a été mesurée grâce au rapport signal/distorsion (le SISDR [17] n'étant pas adapté au calcul par bande) exprimé en décibels, qui est défini pour S_F un signal temporel échantillonné à F et sa reconstruction $\hat{S}_F$ de la manière suivante : $SDR(S_F, \hat{S}_F) = 10 \log_{10} \left(\frac{\|S_F\|^2}{\|S_F - \hat{S}_F\|^2} \right)$.

Une évaluation de la qualité perceptive des reconstructions obtenues avec notre modèle désentrelacé par rapport à la référence DAC a été menée à travers un test MUSHRA [19]. Treize participants ont du noter de 0 (mauvais) à 100 (excellent) la qualité perceptive de quatre versions d'un même extrait audio : une référence, une version encodée par DAC, une version encodée par notre modèle et une ancre (filtre passe bas à 3.5 kHz). Ce test a été effectué pour les reconstructions à 16 kHz et 32 kHz. Deux ensembles de 6 extraits ont été tirés au hasard de l'ensemble de test, pour les deux fréquences d'échantillonnage.

5 Résultats

Les résultats affichés dans la table 1 montrent une légère amélioration des performances de notre modèle à 16 kHz par rapport à la référence, liée à l'étape d'ajustement fin. De même, les mesures de reconstruction observées pour la reconstruction globale à 32 kHz indiquent une amélioration de la qualité de la reconstruction par rapport à la version 32 kHz du modèle DAC.

Les mesures perceptives obtenues par le test MUSHRA et résumées dans la table 2 confirment l'observation que nous

TABLE 3 : Désentrelacement - SDR par bande de fréquences

$\hat{d}_{32kHz}$ vs S_{32kHz}		$\hat{S}_{32kHz}$ vs S_{32kHz}	
$[8 - 16 kHz]$	5.85	$[7.9 - 8.1 kHz]$	5.61
$[0 - 8 kHz]$	3.80	$[0 - 16 kHz]$	5.67

avons faite avec les métriques de reconstruction : pour les reconstructions 16 kHz et 32 kHz, notre modèle est plus performant que la référence, même si les écarts types élevés soulignent que les écarts ne sont pas significatifs.

Le désentrelacement de l'information portée par les tokens extraits de notre codec a été étudié en comparant la sortie de chaque branche ($U(\hat{d}_{16kHz})$ et $\hat{d}_{32kHz}$) au signal d'entrée S_{32kHz} . Les spectrogrammes de ces signaux sont représentés sur la figure 2. La reconstruction issue de la branche 16 kHz n'encode des informations que dans les basses fréquences, car la fréquence d'échantillonnage ne permet pas de reconstruire au-delà de 8 kHz. Quant à la reconstruction de la branche 32 kHz, la plupart des informations sont portées dans les hautes fréquences, même si le désentrelacement doux apporte des corrections dans la bande des basses fréquences. La table 3 rassemble des valeurs de SDR entre S_{32kHz} et $\hat{d}_{32kHz}$, en divisant le calcul entre la bande $[0 - 8 kHz]$ et la bande $[8 - 16 kHz]$. L'information du signal reconstruite par la branche 32 kHz est, en effet, beaucoup plus corrélée à S_{32kHz} dans les hautes fréquences (avec un SDR de 5.85 db), même si certaines informations pertinentes sont également encodées dans les basses fréquences (avec un SDR plus faible de 3.80 db), grâce au désentrelacement doux.

Les résultats de la deuxième partie de la table 3 soulignent également que le SDR calculé dans une bande étroite autour de la fréquence de coupure commune des deux sous-bandes (8 kHz) est presque identique au SDR moyen sur l'ensemble du spectrogramme, indiquant l'absence d'artefacts dans cette zone.

6 Conclusion

Dans ce travail, nous introduisons une approche pour concevoir un codec audio neuronal en incorporant une décomposition fréquentielle des signaux d'entrée, qui facilite le désentrelacement des représentations discrètes. Nous démontrons qu'avec une amélioration de la reconstruction, cette représentation favorise également l'interprétabilité des représentations extraites.

Références

- [1] Stanley Smith STEVENS, John VOLKMANN et Edwin Broomell NEWMAN : A scale for the measurement of the psychological magnitude pitch. *J. Acoust. Soc. Am.*, 8(3):185–190, 1937.
- [2] Diederik P KINGMA et Max WELLING : Auto-encoding variational bayes. *In ICLR*, 2014.
- [3] Aaron Van den OORD, Oriol VINYALS et Koray KAVUKCUOGLU : Neural discrete representation learning. *In NeurIPS*, 2017.
- [4] Alexandre DÉFOSSEZ, Jade COPET, Gabriel SYNNAEVE et Yossi ADI : High fidelity neural audio compression. *Trans. Mach. Learn. Res.*, 2023.
- [5] Rithesh KUMAR, Prem SEETHARAMAN, Alejandro LUEBS, Ishaan KUMAR et Kundan KUMAR : High-fidelity audio compression with improved rvqgan. *In NeurIPS*, 2023.
- [6] Yuhta TAKIDA *et al.* : Hq-vae : Hierarchical discrete representation learning with variational bayes. *Trans. Mach. Learn. Res.*, 2024.
- [7] Chengyi WANG, Sanyuan CHEN, Yu WU, Ziqiang ZHANG, Long ZHOU, Shujie LIU, Zhuo CHEN, Yanqing LIU, Huaming WANG, Jinyu LI *et al.* : Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv :2301.02111*, 2023.
- [8] Ondřej CÍFKA, Alexey OZEROV, Umut ŞİMŞEKLI et Gaël RICHARD : Self-supervised vq-vae for one-shot music style transfer. *In ICASSP*, 2021.
- [9] Naoya TAKAHASHI, Mayank Kumar SINGH et Yuki MITSUFUJI : Hierarchical disentangled representation learning for singing voice conversion. *In Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2021.
- [10] Adam POLYAK *et al.* : Speech resynthesis from discrete disentangled self-supervised representations. *Interspeech Conf.*, 2021.
- [11] Zeqian JU *et al.* : Naturalspeech 3 : Zero-shot speech synthesis with factorized codec and diffusion models. *In ICML*, 2024.
- [12] Xiaoyu BIE, Xubo LIU et Gaël RICHARD : Learning source disentanglement in neural audio codec. *In ICASSP*, 2025.
- [13] Kyle HSU, William DORRELL, James WHITTINGTON, Jiajun WU et Chelsea FINN : Disentanglement via latent quantization. *In NeurIPS*, 2023.
- [14] Yi LUO, Jianwei YU, Hangting CHEN, Rongzhi GU et Chao WENG : Gull : A generative multifunctional audio codec. *arXiv preprint arXiv :2404.04947*, 2024.
- [15] Zafar RAFII, Antoine LIUTKUS, Fabian-Robert STÖTER, Stylianos Ioannis MIMILAKIS et Rachel BITTNER : The musdb18 corpus for music separation. 2017.
- [16] Dmitry BOGDANOV, Minz WON, Philip TOVSTOGAN, Alastair PORTER et Xavier SERRA : The mtg-jamendo dataset for automatic music tagging. *In Int. Conf. Mach. Learn. Workshops (ICML-W)*, 2019.
- [17] Jonathan LE ROUX, Scott WISDOM, Hakan ERDOGAN et John R HERSHEY : Sdr-half-baked or well done ? *In ICASSP*, 2019.
- [18] Michael CHINEN *et al.* : Visqol v3 : An open source production ready objective speech and audio metric. *In Int. Conf. Qual. Multimed. Experience (QoMEX)*, 2020.
- [19] Michael SCHOEFFLER *et al.* : webmushra—a comprehensive framework for web-based listening tests. *J. Open Res. Softw.*, 6(1), 2018.