Les modèles d'ASR peuvent-ils retranscrire les sous-genres de Metal ?

Bastien PASDELOUP[1] Axel MARMORET[1]

[1]IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238, Brest, France

Résumé – Les styles extrêmes de chant associés à la musique Metal sont réputés pour leur intensité, leur saturation vocale et leur faible intelligibilité. Dans cet article, nous étudions la capacité des modèles récents de reconnaissance automatique de la parole (ASR) à retranscrire ce type de voix, en nous plaçant dans un cadre de données hors-distribution (OOD). Pour cela, nous évaluons cinq modèles d'ASR à l'état de l'art sur deux types de corpus : des enregistrements vocaux (chant seul), et un ensemble de chansons extraites de divers sous-genres du Metal. Nous appliquons également des techniques de séparation de sources pour isoler la voix dans les morceaux. Nos résultats montrent que les modèles considérés peinent à retranscrire correctement les paroles des chants extrêmes, particulièrement dans les cas de saturation vocale sévère ou de prosodie atypique.

Abstract – Extreme vocal styles in Metal music are known for their intensity, vocal saturation, and low intelligibility. In this paper, we evaluate the ability of recent Automatic Speech Recognition (ASR) models to transcribe such vocals in an out-of-distribution (OOD) setting. We assess five state-of-the-art ASR models using two types of data: isolated extreme vocal recordings and full metal songs from various subgenres. We also apply source separation techniques to extract vocals from the music tracks. Our results show that these models struggle to accurately transcribe extreme vocals, especially in cases of severe vocal distortion or atypical prosody.

1 Contexte et positionnement

"On ne comprend rien à ce qu'il raconte" est souvent la première réaction d'un(e) auditeur(ice) non familier avec les sous-genres extrêmes du Metal. Certes, mais un modèle de reconnaissance automatique de la parole (ASR) [4] peut-il faire mieux ? Dans cet article, nous évaluons la capacité de modèles récents à transcrire les paroles de chants extrêmes. Nous posons ainsi une double question : *les modèles d'ASR peuvent-ils comprendre ce que disent les chanteur(se)s de Metal ?* ; et, surtout, *les chanteur(se)s de Metal chantent-i(e)ls vraiment ce qui est indiqué dans la pochette de l'album ?*

Les chants Metal sont réputés pour leur saturation vocale et leur faible intelligibilité. Leur spécificité a été démontrée tant du point de vue de la production vocale [3], de l'analyse acoustique [14], et de l'intelligibilité de ces chants [9], cette denière dépendant de l'expérience d'écoute. Récemment, plusieurs travaux ont proposé des approches de classification automatique pour ces styles vocaux [5, 15], notamment à travers le jeu de données EMVD [15], que nous utilisons ici. Pourtant, à notre connaissance, aucun travail n'a encore exploré la transcription automatique de ces styles vocaux.

La reconnaissance automatique de la parole a connu des avancées majeures grâce à l'apprentissage profond et aux architectures de type Transformer [16]. Un enjeu central reste la robustesse des modèles face à des données hors-distribution (OOD – *Out-of-Distribution*) [7]. C'est précisément dans ce cadre que se situe notre étude.

Nous évaluons cinq modèles récents parmi les mieux classés sur le tableau comparatif HuggingFace [1] : `Wav2Vec2.0` [2], `WhisperV2/V3` [12], `Canary` [11] et `Phi-4` [1]. Tous utilisent l'architecture Transformer et possèdent des tailles variant de 317M à 5.6B paramètres.

Nous étudierons ces modèles sur des pistes de chant seules, provenant de *EMVD* [15], puis sur un jeu de données constitué pour cette étude, comprenant des morceaux de musique obtenus via *YouTube* [2]. Notons que `Wav2Vec2.0` et `Canary` nécessitent un ré-échantillonnage à 16kHz, les rendant inadaptés aux chansons complètes sans altération, mais utilisables sur le corpus *EMVD*. Enfin, si la plupart de ces modèles sont multilingues, notre étude se limite à l'anglais, laissant ouverte la question de la performance sur d'autres langues.

L'article est structuré comme suit : les jeux de données sont décrits en Section 2, les expériences et résultats sont fournis en Section 3. Le code, les jeux de données et les résultats sont disponibles à l'adresse : `https://github.com/BastienPasdeloup/extreme_vocals_asr`.

2 Les jeux de données considérés

Une chanson intégrant généralement de nombreux autres instruments que la voix, la tâche de retranscription de parole en est complexifiée. Avant d'évaluer la performance des modèles d'ASR sur de tels morceaux, nous considérons en premier lieu le cas plus simple de pistes audio de chant uniquement.

2.1 Adaptation de *EMVD*

Le jeu de données *Extreme Metal Vocals Dataset (EMVD)* [15] contient des enregistrements de 27 chanteuses et chanteurs interprétant le même texte dans plusieurs styles vocaux. Le corpus présente cependant certaines limitations : les textes n'étaient pas fournis, certains artistes modifiaient les paroles selon le style vocal, et l'un chantait en français. Pour intégrer ce jeu dans notre étude, chaque auteur a fourni

1. `https://huggingface.co/spaces/hf-audio/open_asr_leaderboard`

2. Ces chansons sont les propriétés de leurs ayants droit et sont utilisées ici à des fins strictement scientifiques et non commerciales.

une retranscription indépendante de chaque piste [3] ; lors de l'évaluation, nous retenons la version donnant le meilleur score pour la métrique considérée. Deux artistes (le francophone et un autre impossible à retranscrire) ont été exclus.

Après ce pré-traitement, le corpus final comprend 222 pistes, par 25 chanteurs, réparties selon les catégories suivantes : **Clear Voice** (73), **Hardcore Scream** (51), **Death Growl** (48), **Black Shriek** (42) et **Grind Inhale** (8).

2.2 Les chansons étudiées

Pour évaluer les modèles sur des chansons réelles, tout en essayant d'être représentatifs de la diversité des sous-genres de Metal, nous avons choisi de travailler avec les catégories suivantes : **Black Metal**, **Brutal Death Metal (+ Slam Death)**, **Death Metal**, **Deathcore (+ Metalcore)**, **Doom Metal**, **Depressive Suicidal Black Metal (DSBM)**, **Goregrind**, **Grindcore (+ Deathgrind + Powerviolence)**, **Heavy Metal (+ Power Metal)**, **Melodic Death Metal**, **Punk (+ Hardcore)**, **Symphonic Metal (+ Gothic Metal)**, **Thrash Metal (+ Crossover)**, **War Metal (*a.k.a.* Black/Death)**.

Il est à noter que plusieurs de ces catégories sont des dérivées d'un même sous-genre musical (*e.g.*, **Black Metal**, **DSBM** et **War Metal**). Ce choix se justifie par une esthétique de chant particulière, et pourrait être étendu à des catégories présentant une forte diversité, notamment le Death Metal.

Pour chaque catégorie, 25 morceaux de 25 artistes différents ont été sélectionnés, selon les critères suivants : 1) représentativité dans le sous-genre ; 2) diversité des vocalistes (genre, origine) ; 3) chant en anglais ; et 4) disponibilité en ligne des paroles et des morceaux.

2.3 Séparation de sources

Pour pallier la difficulté liée à la présence d'un accompagnement instrumental dense dans les chansons, nous avons constitué un corpus complémentaire via séparation de sources. Nous avons appliqué le modèle `mdx_extra` de Demucs [13] à l'ensemble des morceaux sélectionnés, afin d'en extraire automatiquement les pistes vocales. Celles-ci sont utilisées en parallèle des versions complètes pour évaluer l'impact de l'accompagnement sur la qualité des transcriptions. Ce corpus intermédiaire permet ainsi de mieux isoler l'effet du contenu vocal dans un contexte réaliste.

3 Expériences

3.1 Approche

Pour chacun des trois jeux de données considérés – *EMVD* (chant seul), *Chansons* (chansons complètes) et *Demucs* (versions vocales des chansons par séparation de sources) – nous générons les transcriptions à l'aide de chacun des modèles présentés en Section 1. La qualité des retranscriptions est évaluée à l'aide des métriques décrites ci-après.

3.2 Métriques

Pour évaluer la qualité des transcriptions, tant syntaxique que sémantique, nous utilisons les métriques suivantes :

- *Word Error Rate (WER)* [8] : mesure classique en ASR, fondée sur le nombre de substitutions, suppressions et insertions nécessaires pour transformer la sortie du modèle en texte de référence. Un score de 0 indique une correspondance parfaite.
- *ROUGE* [6] : évalue le chevauchement lexical (mots ou séquences de mots) entre la prédiction et la référence. Elle fournit une estimation de la fidélité sémantique. Un score de 1 indique une reconstruction parfaite.
- *Autres métriques* : des expériences additionnelles ont été menées avec le score *BLEU* [10], et la similarité cosinus calculée dans des espaces de représentation textuelle fournis par trois modèles de langage. Les observations étant similaires à celles des autres métriques, ces résultats sont uniquement disponibles sur le dépôt GitHub.

3.3 Résultats

La Figure 1 présente les performances obtenues pour chaque jeu de données et chaque métrique. Chaque sous-figure présente les résultats par modèle et par sous-genre.

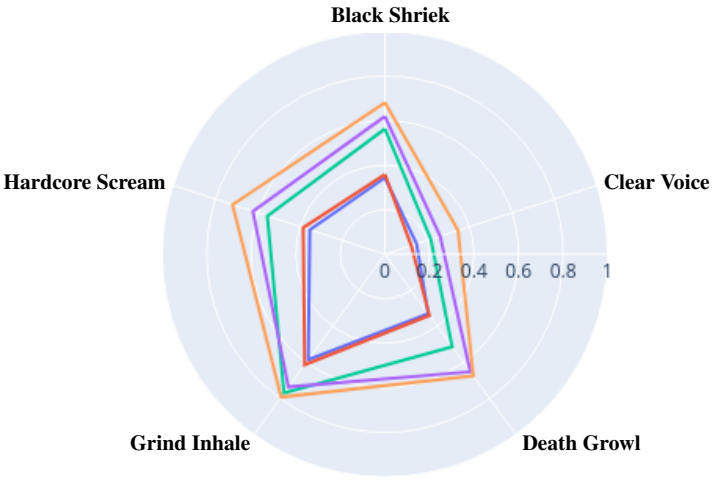
La Figure 2 compare les performances obtenues sur les chansons originales et sur leurs versions après séparation de sources. Seuls les résultats pour `WhisperV3` sont présentés ici, les autres modèles montrant des tendances similaires. Les figures complémentaires sont disponibles sur le dépôt GitHub.

Ces résultats mettent en évidence trois conclusions :

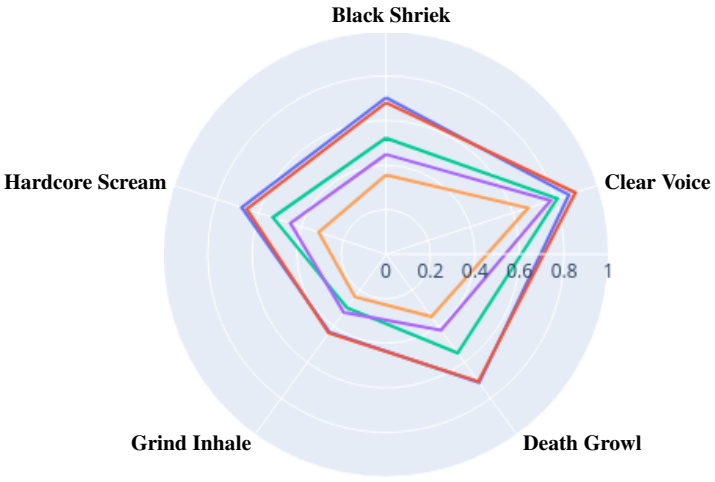
1. Parmi les modèles testés, `WhisperV2/V3` [12] sont les plus performants dans la tâche de transcription du chant Metal, quel que soit le sous-genre.
2. Les performances en transcription varient énormément selon le style de voix. Les performances sont cohérentes avec la perception humaine. En effet, sur les jeux de données *Chansons* et *Demucs*, on distingue quatre groupes : Les chants clairs (**Heavy**, **Symphonic**, **Punk**, **Doom**), les chants moyennement saturés (**Thrash**, **Melodic Death**), les chants saturés (**Black**, **Death**, **Deathcore**) et les chants extrêmes (**War**, **Brutal Death**, **DSBM**, **Grindcore**, **Goregrind**). Ces observations sont cohérentes avec les performances observées dans *EMVD*.
3. La séparation de sources avec Demucs [13] n'a pas montré d'effet significatif sur les performances de transcription. Une explication possible réside dans l'entraînement du modèle `Whisper`, qui inclut des données audio dans des conditions variées de bruit. Cela pourrait permettre au modèle d'apprendre à ignorer les composantes instrumentales pour se concentrer sur la voix. D'ailleurs, les expériences présentées dans l'article original [12] montrent que `Whisper` subit une dégradation moins marquée que d'autres modèles de l'état de l'art face à des perturbations acoustiques, ce qui renforcerait cette hypothèse.

Ces résultats pourront être approfondis par une analyse plus détaillée des performances selon les sous-genres et les styles vocaux, ainsi que par une caractérisation fine des erreurs de transcription (différences voyelles/consonnes, insertions et substitutions dans le WER, impact de la séparation de sources, ou encore ambiguïtés liées à la diction et à l'intelligibilité des paroles dans les enregistrements originaux). Ces investigations sont laissées pour de futurs travaux.

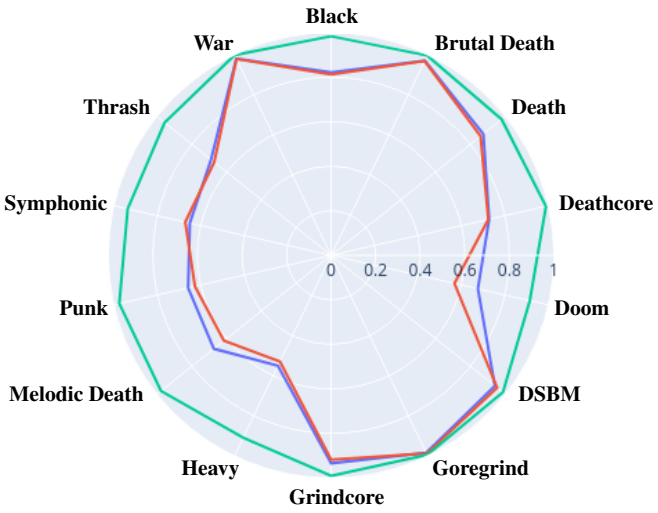
3. Ces retranscriptions ont été transmises aux auteurs de [15].



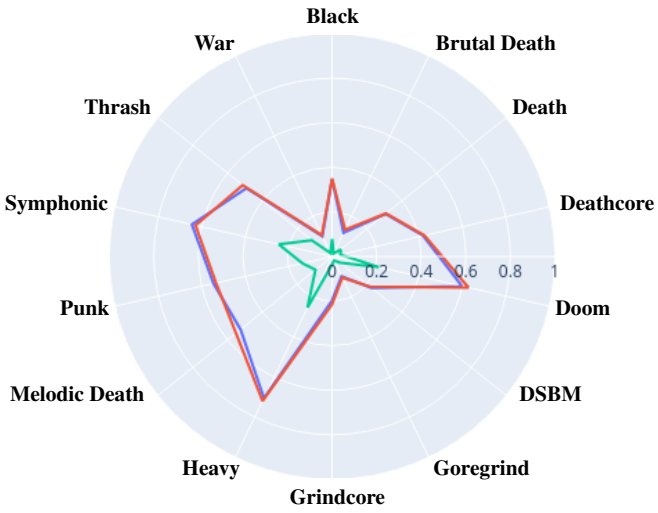
(a) *EMVD* / *WER*



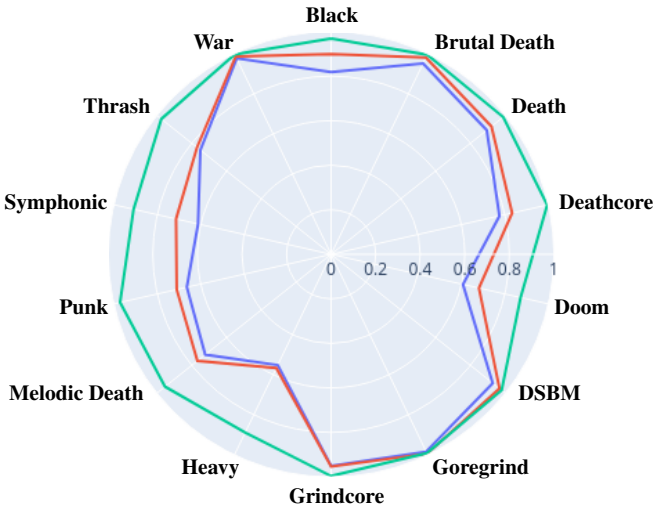
(b) *EMVD* / *ROUGE*



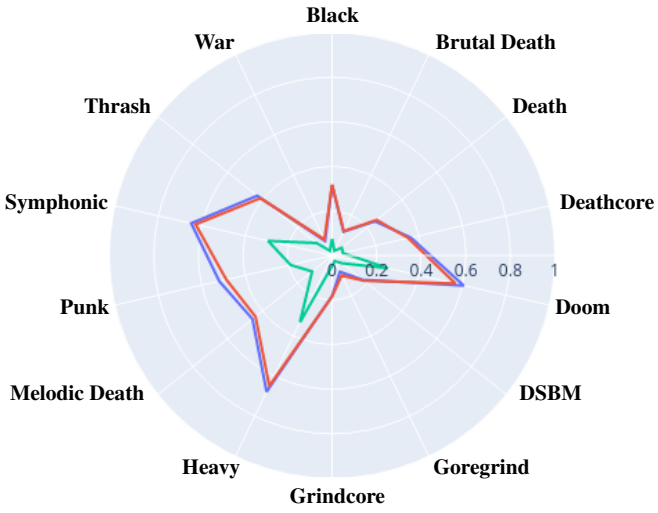
(c) *Chansons* / *WER*



(d) *Chansons* / *ROUGE*

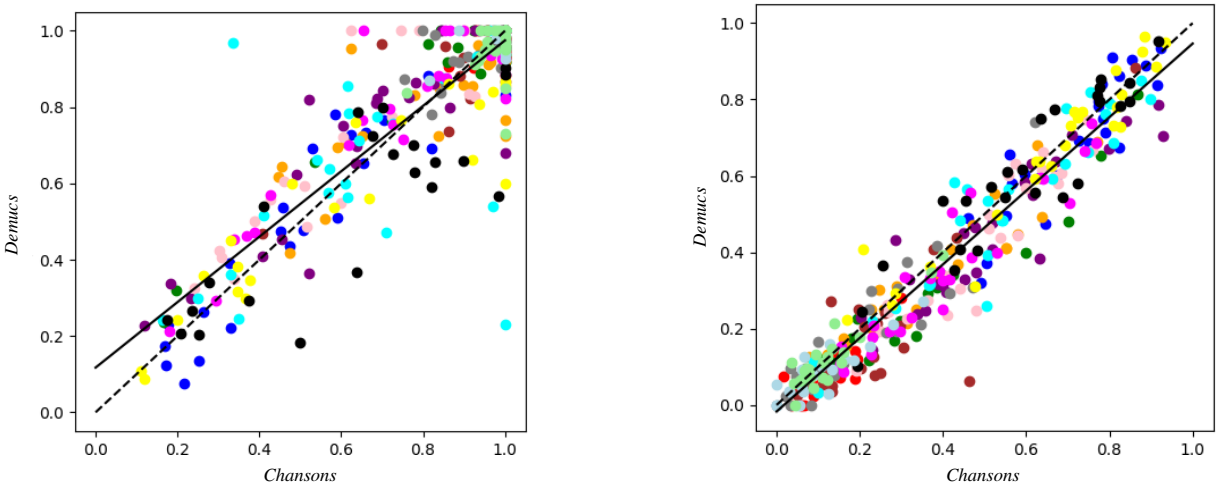


(e) *Demucs* / *WER*



(f) *Demucs* / *ROUGE*

Figure 1 – Performances des modèles `WhisperV3`, `WhisperV2`, `Phi-4`, `Canary` et `Wav2Vec2.0` sur les différents jeux de données et sous-genres, évaluées selon deux métriques : *WER* (plus bas = meilleur) et *ROUGE* (plus haut = meilleur). Le *WER* est seuillé à 1 pour affichage, valeur au delà de laquelle la correspondance est nulle.



(a) *WER* (b) *ROUGE*

Figure 2 – Mise en relation des métriques obtenues pour chaque morceau et sa version source-séparée, avec le modèle `WhisperV3`. Chaque couleur correspond à un sous-genre (**Heavy**, **War**, **Death**, **Black**, **Melodic Death**, **DSBM**, **Thrash**, **Grindcore**, **Punk**, **Deathcore**, **Symphonic**, **Doom**, **Goregrind**, **LimeGreen**). La ligne pointillée correspond à la droite $y = x$, et la ligne continue est une régression linéaire du nuage de points. On observe une corrélation de 0.987 pour *WER* et de 0.953 pour *ROUGE*.

4 Conclusion et perspectives

Nous avons évalué dans cette étude la capacité de cinq modèles récents d'ASR à retranscrire des chants extrêmes, un cas hors distribution (OOD) marqué par une forte saturation vocale et une prosodie atypique. L'analyse, menée sur trois jeux de données complémentaires, montre que si `Whisper` se démarque légèrement, aucun modèle ne parvient à gérer de manière fiable l'ensemble des styles vocaux, en particulier les plus extrêmes (pour lesquels il est relativement convenu pour l'amateur du style que la présence / l'intelligibilité des paroles est secondaire, voire anecdotique). Les résultats varient selon les sous-genres, en accord avec les difficultés perceptives humaines, ce qu'il serait intéressant de confirmer expérimentalement avec un protocole rigoureux. La séparation de sources n'améliore pas significativement les performances, suggérant une certaine robustesse de `Whisper` au bruit. Ces travaux ouvrent des perspectives d'adaptation ciblée des modèles ASR à des voix non conventionnelles. En particulier, dans le cadre de ces travaux, il serait intéressant d'explorer plus finement les erreurs typiques selon le style de chant.

References

[1] Marah Abdin et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[2] Alexei Baevski et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, 2020.

[3] Marco Guzman et al. Aerodynamic characteristics of growl voice and reinforced falsetto in metal singing. *Journal of Voice*, 2019.

[4] Abdelwahab Heba. *Reconnaissance automatique de la parole à large vocabulaire: des approches hybrides aux approches End-to-End*. PhD thesis, Université Paul Sabatier-Toulouse III, 2021.

[5] Vedant Kalbag and Alexander Lerch. Scream detection in heavy metal music. In *Proc. 19th Sound and Music Computing Conf.*, 2022.

[6] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 2004.

[7] Jiashuo Liu et al. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[8] Matteo Negri et al. Quality estimation for automatic speech recognition. In *Proc. 25th Int. Conf. Comput. Linguistics (COLING)*, 2014.

[9] Kirk N Olsen et al. Listener expertise enhances intelligibility of vocalizations in death metal music. *Music Perception: An Interdisciplinary Journal*, 2018.

[10] Kishore Papineni et al. Bleu: a method for automatic evaluation of machine translation. In *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2002.

[11] Krishna C Puvvada et al. Less is more: Accurate speech recognition & translation without web-scale data. In *Proc. Interspeech 2024*, 2024.

[12] Alec Radford et al. Robust speech recognition via large-scale weak supervision. In *Int. Conf. Machine Learning (ICML)*. PMLR, 2023.

[13] Simon Rouard et al. Hybrid transformers for music source separation. In *ICASSP 23*, 2023.

[14] Eric Smialek et al. A spectrographic analysis of vocal techniques in extreme metal for musicological analysis. In *ICMC*, 2012.

[15] Modan Tailleur et al. EMVD dataset: a dataset of extreme vocal distortion techniques used in heavy metal. In *Int. Conf. Content-Based Multimedia Indexing (CBMI)*. IEEE, 2024.

[16] Ashish Vaswani et al. Attention is all you need. In *NeurIPS*, 2017.