Text-aware speech inpainting for restoration of 78rpm disc recordings

Stage de fin d'études d'Ingénieur ou de Master 2 – Année académique 2023-2024

Mots clés : deep learning, large language models, text-to-speech, audio inpainting, audio signal processing

Contexte

L'Institut National de l'Audiovisuel (INA) est un établissement public à caractère industriel et commercial (EPIC), dont la mission principale consiste à sauvegarder et promouvoir le patrimoine audiovisuel français à travers la vente d'archives et la gestion du dépôt légal. Ce dernier permet aux chercheurs en sciences sociales, notamment dans le cadre de projets de recherche pluri-disciplinaires, de travailler sur les collections de l'INA. À ce titre, l'Institut capte en continu 180 chaînes de télévision et radio et stocke plus de 25 millions d'heures de contenu audiovisuel. L'INA assure également des missions de formation, de production et de recherche scientifique.

Le service de la recherche de l'INA mène depuis plus de 20 ans des travaux de recherche dans le domaine de l'indexation et de la description automatique de ces fonds selon l'ensemble des modalités : textes, sons et images. Le service participe à de nombreux projets collaboratifs de recherche que ce soit dans un cadre national et européen et accueille des stages de Master ainsi que des doctorants en co-encadrement avec des laboratoires nationaux d'excellence.

Ce stage est proposé au sein de l'équipe de recherche (https://recherche.ina.fr) et se place au sein de l'équipe INA-Saphir dont l'objectif est d'extraire les sons d'enregistrements sur disques 78 tours très détériorés.

D'autres sujets de stage sont également proposés dans l'équipe : https://www.ina.fr/institut-national-audiovisuel/equipe-recherche/stages

Objectifs du stage

Dans le cadre du projet INA-Saphir, nous avons développé un scanner et des logiciels permettant d'extraire le signal audio de disques 78 tours en très mauvais état. La qualité obtenue reste très souvent mauvaise (souffle et bruits, clicks, interruptions de signal…) et les techniques actuelles de restauration audio restent très insuffisantes. Cependant, les techniques d'apprentissage automatique (Machine Learning, Large Language Models) semblent ouvrir de nouvelles pistes.
Nous désirons au cours du stage explorer la possibilité de reconstruire les voix de tels enregistrements. Nous appuyant sur le timbre, la prosodie et le texte

de l'enregistrement, nous visons à obtenir des pistes de voix re-synthétisées, alignées avec les fragments exploitables, et plausibles sur les parties manquantes. Les parties manquantes ayant des durées de 0.1 à 0.3 secondes, cela suppose de re-construire les mots manquants ou tronqués (Voice Inpainting).
Pour rester dans le cadre d'un stage, nous nous intéresserons ici seulement à évaluer la faisabilité d'une telle reconstruction, connaissant l'intégralité du texte prononcé, ainsi que sa prosodie et son alignement.
Idéalement, une chaine de traitement sera construite, prenant en entrée un enregistrement bruité et sa transcription corrigée, et produisant en sortie une piste audio, alignée sur la source, son timbre, et sa prosodie, non bruitée, et avec les mots de la transcription.
Cette chaine sera probablement constituée d'un ou plusieurs process basés sur le Machine Learning et les LLMs (Large Language Models). Un réapprentissage (fine tuning) sera probablement nécessaire, qui sera effectué sur nos GPUs, sur un corpus d'apprentissage auto-supervisé qui sera également développé lors du stage.
La bibliographie présentée ci-après présente des pistes sur les méthodologies et les outils qui peuvent être envisagés pour le stage.

Valorisation du stage

Différentes stratégies de valorisation des travaux seront envisagées, en fonction de leur degré de maturité et des orientations envisagées pour la suite des travaux :

- Diffusion des outils d'analyse réalisés sous licence open-source via le dépôt GitHub de l'INA : https://github.com/ina-foss
- Rédaction de publications scientifiques

Conditions du stage

Le stage se déroulera sur une période de 4 à 6 mois, au sein du service de la Recherche de l'Ina. Il aura lieu sur le site Bry 2, situé au 28 Avenue des frères Lumière, 94360 Bry-sur-Marne. La·le stagiaire sera encadré·e par Jean-Hugues Chenot (jhchenot@ina.fr). Un ordinateur équipé d'un GPU sera fourni ainsi qu'un accès au cluster de calcul de l'Institut.

Gratification : 760 € brut / mois + 50 % pass Navigo

Candidature

Envoyez par email et en PDF aux adresses jhchenot@ina.fr et ddoukhan@ina.fr les documents suivants : CV et relevés de notes + liste des enseignements suivis en M2 et en M1 ou en école d'ingénieurs. Précisez le sujet du stage pour lequel vous candidatez.

Profil recherché

- Étudiant·e en dernière année d'un bac +5 dans le domaine de l'informatique et de l'IA
- Forte appétence pour la recherche académique ou appliquée
- Maîtrise de Python et expérience dans l'utilisation de bibliothèques de traitement de signal et Machine Learning
- Capacité à effectuer des recherches bibliographiques
- Rigueur, synthèse, autonomie, capacité à travailler en équipe

Bibliographie

[1] Jean-Hugues Chenot, Jean-Étienne Noiré. Challenges in Optical Recovery of Otherwise Unplayable Analogue Audio Disc Records. 2023 AES International Archiving & Preservation Conference, Culpeper (USA). http://www.aes.org/e-lib/download.cfm/22135.pdf?ID=22135

[2] Zalán Borsos, Matt Sharifi, Marco Tagliasacchi "SpeechPainter: Text-conditioned Speech Inpainting". Submitted to Interspeech 2022. https://doi.org/10.48550/arXiv.2202.07273

[3] Pierre Prablanc, Alexey Ozerov, Ngoc Q. K. Duong, Patrick Pérez. "Text-informed speech inpainting via voice conversion." 24th European Signal Processing Conference (EUSIPCO 2016), Aug 2016, Budapest, Hungary. https://inria.hal.science/hal-01271257v2

[4] Andrew Mason "How Imputations Work: The Research Behind Overdub". September 17, 2019. https://www.descript.com/blog/article/how-imputations-work-the-research-behind-overdub

[5] Haohe Liu, Xubo Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, Yuxuan Wang. "VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration" Interspeech 2022, Incheon, Korea. https://cliffzhao.github.io/Publications/LLKTZWHW.INTERSPEECH22.pdf https://github.com/haoheliu/voicefixer

[6] Haohe Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, Yuxuan Wang. "VoiceFixer: Toward General Speech Restoration With Neural Vocoder". 2021. https://arxiv.org/abs/2109.13731.

[7] Wei-Ning Hsu, Tal Remez, Bowen Shi, Jacob Donley, Yossi Adi. "ReVISE: Self-Supervised Speech Resynthesis with Visual Input for Universal and Generalized Speech Regeneration". CVPR2023. https://wnhsu.github.io/ReVISE/ .

[8] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux. "Speech Resynthesis from Discrete Disentangled Self-Supervised Representations". InterSpeech 2021. https://arxiv.org/abs/2104.00355

[9] Sang-Hoon Lee, Eunwoo Song, Seung-Bin Kim, Min-Jae Hwang, Ji-Hyun Lee, Seong-Whan Lee. "HierSpeech: Bridging the Gap between Text and Speech by Hierarchical Variational Inference using Self-supervised Representations for Speech Synthesis". NeurIPS 2022.

[10] He Bai, Renjie Zheng, Junkun Chen, Xintong Li, Mingbo Ma, Liang Huang."A^3T: Alignment-Aware Acoustic and Text Pretraining for Speech Synthesis and Editing". 39 th International Conference on Machine Learning, ICML 2022. https://arxiv.org/abs/2203.09690 https://github.com/richardbaihe/a3t