Projet : "*Voice Separation*"

Cursus : DS Juin 21
Membres: Ephi MALGARAS, Mickaël RIVIER, Stany GALLIER

1. Contexte

La séparation de voix consiste à isoler et extraire la voix humaine d'un fond sonore (musique, bruit,…). En ce sens, elle s'oppose au mixage (où l'on va combiner plusieurs pistes : voix, instrument,...) et est quelquefois qualifiée de « démixage ». La séparation de voix peut avoir de nombreuses applications comme dans les télécommunications (débruitage), l'industrie musicale (karaoké, remasterisation, remixage) ou en reconnaissance vocale, qui est particulièrement sensible au bruit de fond ambiant (le "*cocktail party problem"*). La séparation de voix est ainsi très largement utilisée pour les prothèses auditives notamment et permet d'améliorer nos communications au quotidien (messagerie vocale, assistants vocaux,...).

La séparation de voix s'inscrit dans une problématique plus large qui est celle de la séparation de sources audio et dans laquelle on souhaite reconstruire les propriétés de chacune des composantes du son (séparation en plusieurs instruments/voix, localisation des sources pour l'analyse de scènes auditives, …)

Une conférence vidéo faite par Inria et SONY est disponible en ligne [1] et introduit le contexte et les développements récents dans ce domaine.

2. Etat de l'art

Le problème de séparation de voix a initialement été abordé comme un pur problème de traitement du signal. Au début des années 2010, la méthode la plus convaincante était le multichannel Non-negative Matrix Factorization (NMF) [2] et son Flexible Audio Source Separation Toolbox (FASST). Cependant, les résultats sont mitigés dans la plupart des cas (multiples voix, réverbération). Le domaine de la séparation de voix a plus récemment connu un essor avec le développement et l'application de techniques d'apprentissage profond, ou Deep Learning (DL).

Les premières approches DL datent de 2013 et visaient principalement à débruiter les spectrogrammes obtenus par Short-Time Fourier Transform (STFT), généralement présentés en échelle Mel [13]. Des masques robustes dans l'espace temps-fréquence permettent alors d'isoler la voix du reste de l'audio [3,4]. Par la suite, des méthodes de *Deep Clustering*, reposant sur une architecture encodeuse pour réduire la dimension du problème de clustering, ont permis d'améliorer la qualité de la séparation, et de se généraliser avec succès à un problème à trois voix [5].

Un point central dans l'utilisation des techniques de Deep Learning est le choix de la fonction coût (ou loss function). Ainsi, différents problèmes, comme le démixage ou le *cocktail party problem*, peuvent être traités de manière similaire en utilisant des fonctions coût adaptées. Par exemple, l''utilisation de fonctions coût invariantes à la permutation, Permutation Invariant Training (PIT), permet une sensible amélioration des résultats dans le contexte du *cocktail party problem* [6].

Ces approches dans l'espace fréquentiel présentent cependant quelques limitations (*e.g.* utilisation limitée du phasage), et les travaux plus récents tendent à se concentrer sur l'espace temporel uniquement. L'approche Conv-TasNet [7], proposée en 2019, a ainsi grandement dépassé les résultats obtenus grâce au Deep Clustering. Cette approche conserve l'architecture générale classique Encodeur, Séparateur, Décodeur. Cependant, elle apporte de nombreuses améliorations, que ce soit dans les phases d'encodage/décodage (remplacement de la STFT et des masques par des convolutions entraînées), ou dans la phase de séparation, réalisée par un Temporal Convolutional Network (TCN).

En pratique, de nombreux défis subsistent encore dans le domaine de la séparation de voix. On peut citer par exemple la séparation de voix ou d'instruments dont on ne connaît pas le nombre a priori [8], la gestion des données bruitées ou faiblement labellisées, la classification et description des sons ou encore l'amélioration et le débruitage de voix en temps réel.

Des articles d'overview sont disponibles sur le domaine du démixage [9], du cocktail party problem [10], ou plus généralement sur l'utilisation des méthodes DL dans la séparation de voix [11]

Enfin, les codes de différents articles sont disponibles sur le site Papers with code [14], permettant une comparaison objective des méthodes. Les résultats obtenus aux diverses compétitions telles que SiSEC2021, InterSpeech2020 et ICASSP2021 permettent eux aussi d'avoir une vision d'ensemble sur les méthodes à l'état de l'art.

3. Objectifs du projet proposé

Dans le cadre de ce sujet, nous proposons de débuter par le problème le plus simple de la séparation de voix d'un fichier audio musical. L'idée est donc à partir d'un fichier audio (musique et voix) d'en extraire la partie vocale seule. Au vu des datasets disponibles pour commencer l'étude, nous nous concentrerons dans un premier temps sur le problème du démixage, avec comme objectif de reproduire les étapes importantes des développements récents (NMF, masque en temps-fréquence, approches convolutionnelles, …). Nous nous comparerons à l'état de l'art grâce à des outils open sources comme Open-Unmix.

Par la suite, selon l'avancement, nous pourrons nous pencher sur le *cocktail party problem* (avec par exemple l'outil Asteroid de PyTorch), ou bien proposer une application concrète de l'outil de démixage, comme la traduction des paroles d'une chanson, ou encore la détection de présence de chaque instrument en temps réel.

4. Bases de données disponibles

- MUSDB18 : 150 morceaux de musiques (~10h) avec les channels isolés suivants : "drums", "bass", "vocals" et "others".
- DSD100 : 100 morceaux de musiques avec les channels isolés suivants : "drums", "bass", "vocals" et "others".
- DALI : 5358 pistes audio avec l'alignement des paroles connues.

D'autres datasets sont disponibles en ligne, orientés vers le démixage ou le *cocktail party problem*, tels que MSD100, wsj0-mix2, WildMix, FUSS et Slakh2100. Attention, les trois

derniers sont non-supervisés, on ne connaît donc pas les sources initiales ayant menées aux mixtures disponibles.

5. Difficultés - Points de vigilance

Comme abordé dans l'état de l'art, la fonction coût joue un rôle prépondérant dans la qualité des résultats des méthodes DL et doit être choisie avec soin selon l'objectif de l'étude.

Un point également important est la métrique d'évaluation [12], permettant de juger de la qualité de la séparation. Souvent de type SI-SNR ou SI-SDR (Scale-Invariant Signal-to-Noise Ratio/Source-to-Distortion Ratio). Une bon score n'implique cependant pas toujours un bon rendu du fait de la subjectivité de la perception humaine

6. Outils complémentaires

En plus des outils de DL, il sera nécessaire d'utiliser des outils de traitement audio/ traitement du signal. On peut ainsi citer les librairies : Librosa, Scipy_mmap, tensorflow io_fromaudio ou encore torchaudio.

Des packages Python sont également généralement associés aux bases de données (ex. Museval pour la base MUSDB18, elle permet notamment de calculer les métriques abordées précédemment).

Finalement, les principaux outils open-source, avec les meilleurs résultats, sont les suivants : Open-Unmix, Demucs, Spleeter et Nussl pour le démixage ; Asteroid pour le *cocktail party problem*.

7. Bibliographie

Vidéo conférence

1. IAES Virtual Symposium: Applications of Machine Learning in Audio September 28-29 2020. Speaker: Fabian-Robert Stöter (Inria), Stefan Uhlich (Sony) https://youtu.be/AB-F2JmI9U4

Articles de journaux

2. Ozerov, A., Févotte, C., & Vincent, E. (2018). An introduction to multichannel nmf for audio source separation. In *Audio Source Separation* (pp. 73-94). Springer, Cham.
3. Narayanan, A., & Wang, D. (2013, May). Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 7092-7096). IEEE.
4. Huang, P. S., Kim, M., Hasegawa-Johnson, M., & Smaragdis, P. (2014, May). Deep learning for monaural speech separation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1562-1566). IEEE.
5. Hershey, J. R., Chen, Z., Le Roux, J., & Watanabe, S. (2016, March). Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 31-35). IEEE.
6. Yu, D., Kolbæk, M., Tan, Z. H., & Jensen, J. (2017, March). Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017*

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 241-245). IEEE.
7. Luo, Y., & Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, *27*(8), 1256-1266.
8. Nachmani, E., Adi, Y., & Wolf, L. (2020, November). Voice separation with an unknown number of multiple speakers. In *International Conference on Machine Learning* (pp. 7164-7175). PMLR.
9. Rafii, Z., Liutkus, A., Stöter, F. R., Mimilakis, S. I., FitzGerald, D., & Pardo, B. (2018). An overview of lead and accompaniment separation in music. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(8), 1307-1335.
10. Gannot, S., Vincent, E., Markovich-Golan, S., & Ozerov, A. (2017). A consolidated perspective on multimicrophone speech enhancement and source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(4), 692-730.
11. Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1702-1726.
12. Vincent, E., Gribonval, R., & Févotte, C. (2006). Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, *14*(4), 1462-1469.

Ressources en ligne

13. Getting to know the Mel spectrogram, *Towards Data Science* https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0
14. Speech separation, *Papers with code* https://paperswithcode.com/task/speech-separation