Rapport de stage

Master 2 Automatique Robotique
Parcours: Ingénierie des Systèmes Intelligents

Transcription Automatique avec CNN pour la cithare Malgache *marovany*

Rahaoui Linda

Encadrants :
Marc Chemillier
Olivier Adam

Année 2023/2024

Contents

Résumé du stage **3**
0.1 Résumé . . . 3
0.2 Abstract . . . 3

1 Introduction **5**

2 Présentation synthétique du laboratoire **7**

3 Contexte et Enjeux **10**
3.1 Contexte . . . 10
3.2 Enjeux . . . 10
3.3 La cithare *marovany* . . . 10
3.4 Missions . . . 12

4 La Transcription Automatique **13**
4.1 Multi-Pitch Estimation (MPE) . . . 13
4.2 Note Tracking . . . 13
4.3 Méthodes Avancées : NMF et Réseaux de Neurones . . . 14
4.3.1 Factorisation en Matrice Non-Négative (NMF) . . . 14
4.3.2 Réseaux de Neurones . . . 15
4.4 Méthodes de Pointe : L'Approche de Google Brain . . . 15

5 Présentation de la méthode de transcription **17**
5.1 Basic Pitch . . . 17
5.2 CREPE notes . . . 17
5.3 Amélioration de CREPE Notes . . . 19
5.4 Ré-entrainement du CNN de Madmom . . . 20
5.4.1 Comparaison des Architectures . . . 21

6 Expérimentation **22**
6.1 La Dataset . . . 22
6.1.1 Segmentation et Annotation Semi-Automatique . . . 22
6.1.2 Génération de Données Synthétiques . . . 23
6.1.3 Ajout de Bruits pour Diversification . . . 23
6.2 Pré-traitements des données . . . 24
6.3 Entraînement . . . 25

7 Résultats **27**
7.1 Évaluation 27
7.2 Résultats 27

8 Conclusion **30**

Bibliographie **31**

0.1 Résumé

Ce mémoire porte sur la transcription automatique de la musique jouée sur la cithare *marovany*, un instrument traditionnel malgache complexe, dont la transmission se fait principalement de manière orale. L'objectif de cette recherche est d'exploiter des techniques modernes d'intelligence artificielle, notamment les réseaux de neurones convolutionnels (CNN), afin de développer des transcriptions robustes pour cet instrument, contribuant ainsi à la préservation de ce patrimoine musical à travers la création de documentations.

Après une revue des méthodes actuelles de transcription automatique, ce travail se focalise sur l'amélioration d'une méthode spécifique, CREPE Notes, qui utilise deux CNN pour la détection du pitch et des attaques de notes, associée à une approche de segmentation des contours de pitch. Parmi les innovations introduites, on relève l'augmentation de données par la création de données synthétiques, l'amélioration d'architectures de réseaux de neurones convolutionnels pour une meilleure détection des notes, ainsi que l'application de techniques de filtrage et de post-traitement des transcriptions.

Ces améliorations ont permis d'accroître significativement la précision des transcriptions par rapport aux méthodes de référence, démontrant ainsi l'efficacité des techniques avancées d'intelligence artificielle appliquées à la transcription automatique de musique traditionnelle. L'étude met également en lumière l'importance cruciale d'une base de données bien construite pour la transcription automatique, notamment pour des instruments aussi spécifiques que la cithare *marovany*. Enfin, ce mémoire propose plusieurs pistes d'amélioration, notamment la création d'une base de données plus représentative du jeu de la cithare et l'exploration de modèles plus complexes pour accroître encore la précision des transcriptions futures.

0.2 Abstract

This thesis focuses on the automatic transcription of music played on the *marovany* zither, a complex traditional Malagasy instrument, whose transmission is primarily oral. The objective of this research is to leverage modern artificial intelligence techniques, particularly convolutional neural networks (CNNs), to develop robust transcriptions for this instrument, thereby contributing to the preservation of this musical heritage through the creation of documentation.

Following a review of current automatic transcription methods, this work focuses on improving a specific method, CREPE Notes, which utilizes two CNNs for pitch detection and note onset detection, combined with a pitch contour segmentation approach. Among the innovations introduced are data augmentation through the creation of synthetic data, improvements to convolutional neural network architectures for better note detection, and the application of filtering and post-processing techniques to transcriptions.

These improvements have significantly increased the accuracy of transcriptions compared to reference methods, demonstrating the effectiveness of advanced artificial intelligence techniques applied to the automatic transcription of traditional music. The study

also highlights the crucial importance of a well-constructed database for automatic transcription, especially for such specific instruments as the *marovany* zither. Finally, this thesis proposes several avenues for improvement, including the creation of a more representative database of *marovany* zither music and the exploration of more complex models, such as those combining CNNs and LSTMs, to further enhance the accuracy of future transcriptions.

Chapter 1

Introduction

La Transcription Automatique de Musique (Automatic Music Transcription, AMT) est une méthode, appartenant au domaine de la recherche d'information musicale (Music Information Retrieval, MIR). Elle consiste à convertir des enregistrements musicaux en une représentation symbolique, par exemple sous forme de fichier MIDI. Cette technique combine le traitement du signal et l'intelligence artificielle pour extraire et décomposer les caractéristiques complexes des enregistrements, les traduisant en données musicales simplifiées et exploitables. L'AMT se décompose généralement en deux sous-tâches : l'estimation des hauteurs des notes (multi-pitch estimation) et la détection des débuts et fins des notes (onsets et offsets detection). Pour obtenir des transcriptions précises, d'autres informations telles que le tempo et l'intensité sont également extraites.

L'AMT reste aujourd'hui un défi difficile et non résolu dans le domaine du MIR en raison de la complexité de certains morceaux de musique et de certains intruments. Notamment, les morceaux composés de plusieurs notes jouées simultanément ou encore composés de plusieurs sources sonores (chant, instruments).

Lors de mon stage de première année de Master, j'ai travaillé sur la création d'un algorithme de transcription automatique pour un instrument traditionnel malgache, la cithare *marovany*. Cet instrument, central dans la culture malgache, est utilisé lors d'événements importants tels que des célébrations, des cérémonies et des rituels religieux. La musique de la *marovany* est traditionnellement transmise oralement de génération en génération et n'est pas retranscrite par écrit, ce qui pose des problèmes de préservation pour les futures générations.

Mon travail a consisté à combler ce manque de documentation en développant une application spécifiquement dédiée à la *marovany*. Cette application, basée sur des techniques de traitement du signal, était capable de générer des fichiers MIDI à partir d'enregistrements audio au format WAV. Pour pallier les limites des méthodes de traitement du signal, j'ai également intégré des outils permettant de modifier manuellement les transcriptions. Bien que cette première approche ait produit des résultats encourageants, elle nécessitait encore de nombreux ajustements manuels, en particulier pour les enregistrements longs, rendant le processus fastidieux.

C'est pourquoi, pour mon stage de deuxième année de Master, l'objectif était d'améliorer la méthode de transcription de la cithare en utilisant des techniques de machine learning,

et plus particulièrement les réseaux de neurones. Cette nouvelle approche visait à rendre l'algorithme plus robuste et autonome, en réduisant au minimum les ajustements manuels nécessaires pour obtenir des transcriptions précises.

Nous commencerons par présenter le cadre et les défis qui ont motivé ce projet. Ensuite, nous examinerons les différentes méthodes et techniques existantes, ainsi que celles que j'ai mises en œuvre pour atteindre les objectifs du stage. Enfin, nous analyserons les résultats obtenus, les comparerons avec ceux du stage précédent, et discuterons des difficultés rencontrées au cours du projet.

Chapter 2

Présentation synthétique du laboratoire

L'Institut

L'Institut Jean le Rond d'Alembert est un centre de recherche de pointe dédié à l'étude de l'acoustique, de la mécanique et de l'énergétique. Affilié à Sorbonne Université (anciennement UPMC), au CNRS et au Ministère de la Culture, ce laboratoire rassemble des experts de plusieurs disciplines au sein de cinq divisions principales :

- Fluides Complexes et Instabilités Hydrodynamiques (FCIH)
- Combustion Énergie Propre Turbulence (CEPT)
- Lutheries - Acoustique - Musique (LAM)
- Mécanique et Ingénierie des Solides et des Structures (MISES)
- Modélisation, Propagation et Imagerie Acoustique (MPIA)

L'équipe du LAM

Mon stage s'est déroulé au sein de l'équipe Lutheries - Acoustique - Musique (LAM), une unité de recherche interdisciplinaire qui se concentre sur l'étude détaillée des phénomènes sonores et musicaux. Le LAM aborde le son et la musique sous trois angles complémentaires : les sciences de l'ingénierie (notamment en physique et acoustique), les sciences humaines (telles que la musicologie et la psychologie cognitive), et la recherche musicale (les orchestres électro-acoustiques, les performances musicales, et les instruments augmentés).

Le groupe est composé de chercheurs, enseignants-chercheurs et doctorants issus de diverses disciplines, tous animés par une passion commune pour la musique. Leur travail se divise en trois axes de recherche principaux :

1. **L'instrument de musique en situation** : Cet axe explore les caractéristiques sonores produites par les instruments de musique à travers leurs structures vibratoires et vibro-acoustiques, en se concentrant sur leur conception et leur fonctionnement. Les compétences en acoustique, vibration, traitement du signal, psychoacoustique, aéroacoustique et mécanique des fluides sont mobilisées pour analyser

le fonctionnement interne des instruments, la diffusion du son et la perception globale qui en résulte. Ce travail englobe une large gamme d'instruments tels que les violons, guitares électriques, flûtes, orgues, et voix chantées. Des collaborations internationales avec des institutions académiques et musicales permettent d'enrichir ces recherches.

2. **Les nouvelles lutheries et la voix** : Les travaux de cette section se concentrent sur la performance et la création musicale, en explorant de nouvelles approches pour la conception et la perception sonore à travers des instruments numériques et des interfaces Homme-Machine. Les chercheurs s'intéressent à des domaines variés tels que la voix, la synthèse vocale, et le développement de dispositifs sonores et vibrotactiles. Les interfaces vibrotactiles, en particulier, ouvrent la voie à des applications novatrices dans les domaines thérapeutiques et de communication.

3. **Les espaces sonores** : Cet axe se divise en trois sous-axes : l'acoustique en réalité virtuelle et augmentée, l'acoustique architecturale et archéologique, et la perception sonore spatiale. L'acoustique architecturale et archéologique explore la propagation, la résonance et l'interaction du son avec l'espace. L'étude de l'audio spatial vise à mieux comprendre comment les sons sont localisés, interprétés et perçus dans notre environnement. Enfin, les recherches en réalité virtuelle, réalité augmentée et interfaces audio s'intéressent à l'impact potentiel du son 3D dans diverses applications, telles que les outils de simulation, les dispositifs d'assistance, et les rendus audio-visuels.

L'équipe de M.Chemillier

Le CAMS (Centre d'analyse et de mathématique sociales, unité mixte EHESS – CNRS UMR 8557) est une unité résolument pluridisciplinaire : ses chercheurs et enseignants-chercheurs sont issus de champs disciplinaires variés (mathématiques, informatique, physique, économie, anthropologie, géographie...), et ont acquis et développent des compétences aux interfaces. Le CAMS occupe une place originale dans le paysage de la recherche en France, à l'articulation entre sciences sociales d'une part, et mathématiques, informatique, et physique théorique d'autre part. Comme centre de l'EHESS et comme unité mixte du CNRS, le CAMS porte une double ambition : celle de participer à la recherche mathématique internationale au meilleur niveau dans les domaines de compétence où il a acquis une notoriété, et celle de remplir pleinement son rôle d'innovation, de diffusion et de passage à la fois de la mathématique et de l'informatique vers les sciences sociales et, en sens inverse, de problématisation mathématique originale de questions venant des sciences sociales et, plus largement, des sciences du vivant.

Au sein du CAMS, Marc Chemillier mène des travaux sur la modélisation des savoirs relevant de l'oralité, et plus particulièrement sur les savoirs musicaux. Ils se situent à la croisée de l'anthropologie, de la musicologie, de l'informatique et des mathématiques. La modélisation consiste à étudier les principes de cohérence qui organisent ces savoirs et à en explorer la logique sous-jacente. Dans ce cadre, il s'intéresse en collaboration avec l'IRCAM (Institut de Recherche et de Coordination Acoustique/Musique) à la modélisation des savoirs mis en jeux dans l'improvisation musicale en jazz et dans les musiques traditionnelles, notamment à Madagascar, et plus particulièrement aux mutations provoquées dans la création artistique par l'utilisation généralisée des nouvelles technologies.

Sur le plan mathématique, ses travaux de modélisation musicale se rattachent au domaine de la combinatoire des mots et de la théorie des langages formels.

Ces recherches ont été soutenues récemment par le projet européen ERC REACH 2020-2025 (Raising Co-Creativity in Cyber-Human Musicianship), qui a donné naissance à plusieurs intelligences artificielles pour l'improvisation musicale parmi lesquelles le logiciel Djazz utilisée par Marc Chemillier pour la musique de Madagascar. La particularité de ce logiciel est de capter le jeu d'un musicien en rythme avec une pulsation régulière et de faire un apprentissage sur ces données pour pouvoir ensuite improviser à la place du musicien tout en suivant le rythme. Marc Chemillier s'intéresse en particulier au jeu de la cithare malgache *marovany* très virtuose sur le plan rythmique. Les données MIDI obtenues par le système de transcription automatique décrit dans le présent mémoire sont destinées à alimenter l'IA Djazz. Marc Chemillier a déjà donné de nombreux concert en duo avec des musiciens malgaches en utilisant cette IA (Chemillier and Vali 2022, Chemillier 2025), notamment le grand cithariste Justin Vali qui s'est fait connaître par ses collaborations avec Kate Bush et Peter Gabriel. Il a utilisé pour cela des données MIDI issues d'un précédent système de détection automatique sur la cithare *marovany* (Cazau et al. 2016). Le stage de M2 décrit dans ce mémoire a été financé par le projet ERC REACH.

Chapter 3

Contexte et Enjeux

3.1 Contexte

Pour mon stage de première année, le projet de transcription automatique de la cithare s'appuyait sur les travaux de thèse de Dorian Cazau. Cette thèse, motivée par le projet projet ANR IMPROTECH 2010-2013 du professeur Marc Chemillier, proposait une approche de transcription musicale en intégrant des connaissances préalables sur le timbre et les techniques de jeu de la *marovany*. L'approche de M. Cazau reposait sur l'utilisation de méthodes statistiques telles que le PLCA (Probabilistic Latent Component Analysis) et les HMM (Hidden Markov Models) pour estimer les hauteurs, la durée, et la structure temporelle des notes. Ces recherches visaient à créer un système de transcription dédié à la cithare *marovany*.

3.2 Enjeux

La complexité du jeu de cet instrument traditionnel rend sa transcription difficile. La musique de la cithare, qui reflète les traditions, la culture et les coutumes des Malgaches, est transmise presque exclusivement de manière orale. L'absence de documentation et de méthodes modernes pour transmettre ces traditions aux nouvelles générations entraîne une perte progressive de cet héritage. Un algorithme spécifiquement conçu pour la cithare pourrait offrir la possibilité de créer des transcriptions précises et d'obtenir des partitions de musique. Ces partitions permettraient de créer une documentation pour préserver et transmettre la musique de la *marovany* ainsi que les traditions malgaches. De plus, elles pourraient servir de données d'entrée à un logiciel de co-improvisation conçu par une équipe supervisée par M. Chemillier dans le cadre du projet ANR MERCI 2019-2023 et du projet ERC REACH 2020-2024. Cela offrirait la possibilité de faire découvrir cette musique traditionnelle en jouant avec d'autres musiciens du même registre via un ordinateur ou encore un synthétiseur.

3.3 La cithare *marovany*

La cithare *marovany* est fabriquée à partir d'un cadre rectangulaire en bois et de cordes métalliques recyclées. Les méthodes et les matériaux utilisés pour fabriquer cet instru-

ment varient, permettant aux musiciens de personnaliser leur cithare en choisissant le type de bois, les cordes et leurs dimensions. Souvent, les câbles de frein de motos, d'environ 1 mètre de long, sont utilisés comme cordes. La cithare est généralement équipée de 20 à 24 cordes, réparties de chaque côté du cadre en bois et fixées sur des chevalets. Jouée en pinçant ou en grattant les cordes avec les doigts ou les ongles, la *marovany* permet, grâce à son grand nombre de cordes, de créer des mélodies variées, riches et complexes.

Figure 3.1: La cithare *marovany* utilisée pendant le stage.

Le style musical joué par les citharistes se caractérise par des motifs répétitifs et rapides. Cette rapidité entraîne une résonance prolongée de certaines cordes, tandis que d'autres vibrent sans avoir été directement jouées, générant ainsi une "résonance par sympathie" qui enrichit la texture de la musique traditionnelle.

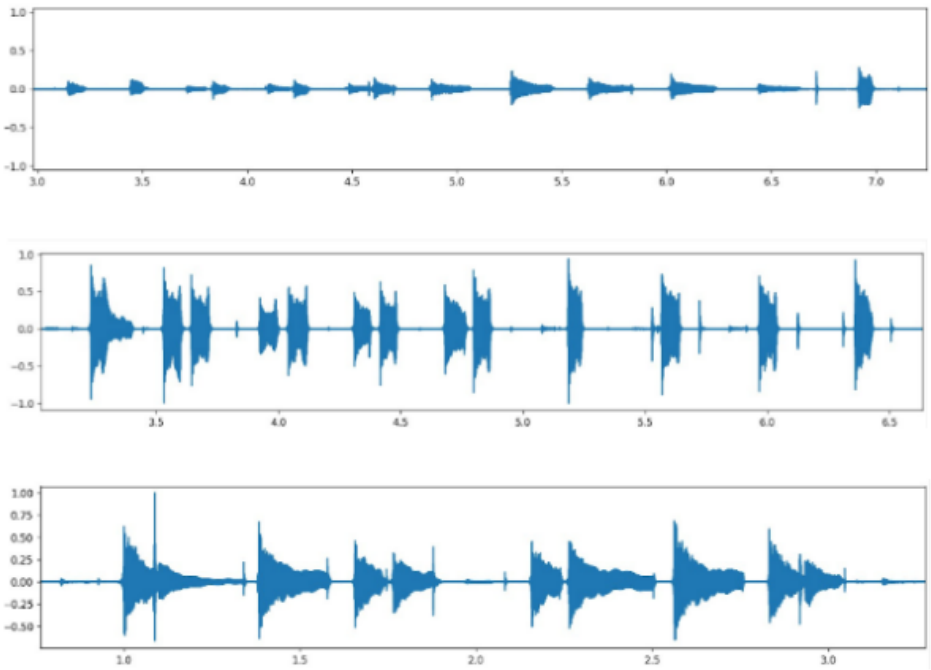

Figure 3.2: Exemple de différents signaux de jeux de cithare.

La cithare utilisée pendant mon stage (cf. Figure 3.1) est équipée de 12 cordes de chaque côté. Des capteurs piézoélectriques ont été placés sur chacune des cordes pour capturer les signaux audio par vibration, individuellement pour chaque corde. Cette approche simplifie grandement la tâche de transcription, car elle permet de décomposer le signal polyphonique de la cithare en plusieurs signaux monophoniques distincts [42].

3.4 Missions

Mon stage a pour objectif de proposer une solution intégrant des réseaux de neurones, afin d'obtenir une meilleure transcription que celle basée sur le traitement du signal (stage Master 1), tout en limitant au maximum la nécessité d'effectuer des corrections manuelles.

Chapter 4

La Transcription Automatique

Les recherches sur la Transcription Automatique de Musique (AMT) sont en constante évolution depuis plusieurs décennies. Son principal objectif est de fournir une représentation simplifiée de la musique en capturant les caractéristiques essentielles des signaux acoustiques. Par exemple, dans le cadre de la transcription d'instruments, cette tâche se concentre sur l'extraction de trois éléments clés pour chaque note musicale : le moment où elle commence (onset), le moment où elle se termine (offset), et la hauteur à laquelle elle est jouée (pitch) [1]. Ces trois paramètres permettent de caractériser précisément chaque note.

La transcription peut être décomposée en deux sous-tâches fondamentales : l'estimation des hauteurs (Multi-Pitch Estimation, MPE) et la détection des onsets et offsets des notes (note-tracking).

4.1 Multi-Pitch Estimation (MPE)

L'estimation multi-pitch consiste à déterminer le nombre et les hauteurs des notes présentes simultanément dans chaque séquence d'un signal acoustique. Plusieurs méthodes ont été développées pour cette tâche, allant des techniques basées sur le traitement classique du signal [17, 26, 32], aux méthodes utilisant la factorisation en matrices non négatives (NMF) [20,30,31], en passant par des approches exploitant les réseaux de neurones [9,35]. Les méthodes traditionnelles de traitement du signal, bien que rapides et généralisables à différents instruments, manquent parfois de précision face aux approches plus récentes. En revanche, les réseaux de neurones profonds, bien que plus coûteux en calcul, offrent de meilleures performances, notamment pour des instruments complexes tels que la cithare *marovany*.

4.2 Note Tracking

Le note tracking consiste à reconstruire les notes musicales complètes en utilisant les estimations de hauteur (pitch) pour identifier non seulement la hauteur des notes, mais aussi les moments où elles commencent (onsets) et se terminent (offsets) [2]. Plusieurs stratégies peuvent être employées : l'estimation des pitches suivie de la détection des onsets, la détection des onsets suivie de l'estimation des pitches, ou l'estimation simultanée des onsets et offsets avec le même algorithme [3, 24]. Certaines méthodes détectent directement les onsets à partir du signal acoustique [29], tandis que d'autres appliquent

un post-traitement aux résultats de l'estimation de pitch obtenus via une méthode de MPE pour retrouver les onsets et offsets [34]. Parmi les techniques utilisées pour le note tracking, on trouve le filtrage médian [16], les modèles de Markov cachés (HMM) [28], les réseaux de neurones [29], ainsi que les modèles de probabilité spectrale [1] et les modèles de langage musical [27].

4.3 Méthodes Avancées : NMF et Réseaux de Neurones

4.3.1 Factorisation en Matrice Non-Négative (NMF)

La NMF est une méthode qui décompose un spectrogramme en deux matrices : un dictionnaire de modèles spectraux **W** et une matrice d'activations temporelles **H**. La factorisation est exprimée mathématiquement comme suit :

$$V \approx WH \tag{4.1}$$

où **V** est la matrice des spectrogrammes d'entrée, **W** est la matrice dictionnaire, et **H** représente la matrice des activations [39].

Chaque colonne du dictionnaire **W** représente le spectre d'un son particulier, tandis que les activations temporelles dans **H** indiquent quand et à quelle intensité ces sons se produisent. Cette technique permet de décomposer un signal complexe en composants élémentaires et interprétables, comme illustré dans la figure 4.1.

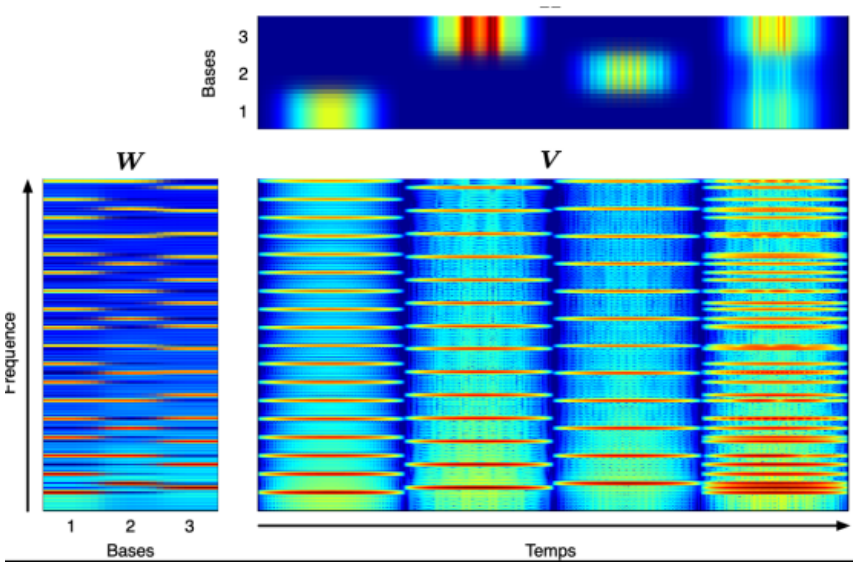

Figure 4.1: Illustration de la Factorisation en Matrice Non-Négative (NMF). **V** est le spectrogramme d'entrée, **W** est le dictionnaire de bases spectrales, et **H** est la matrice des activations temporelles. [39]

Cependant, les modèles NMF sont limités lorsqu'il s'agit de gérer la forte corrélation entre les sons musicaux, notamment lorsque les harmoniques des différentes notes se chevauchent, rendant la séparation difficile [2].

4.3.2 Réseaux de Neurones

Avec les avancées continues du machine learning, les réseaux de neurones ont pris une place prépondérante dans l'AMT. Ils sont particulièrement efficaces pour traiter des données séquentielles ou extraire des caractéristiques complexes des signaux audio. Parmi les types de réseaux les plus couramment utilisés pour l'AMT, on trouve :

- **Réseaux de Neurones Convolutionnels (CNN)** [8,9,33] : Les CNN sont largement utilisés pour l'extraction de caractéristiques à partir de données spectrales, telles que les spectrogrammes ou les transformations en constante Q (CQT). Ils capturent les motifs spatiaux dans les données audio, ce qui les rend efficaces pour identifier les hauteurs de notes et les onsets dans les enregistrements musicaux.

- **Réseaux de Neurones Récurrents (RNN)** [10, 11] : Les RNN, en particulier les variantes LSTM (Long Short-Term Memory) et GRU (Gated Recurrent Unit), capturent les dépendances temporelles dans les signaux audio. Ils sont cruciaux pour modéliser la séquence des notes dans le temps, une tâche essentielle pour la transcription de musique polyphonique.

- **Réseaux de Neurones à Attention (Transformers)** [12, 33] : Les modèles basés sur les Transformers gagnent en popularité pour l'AMT en raison de leur capacité à capturer des relations à long terme dans les séquences de données, sans les limitations des RNN en termes de dépendance à la position dans la séquence.

- **Modèles Hybrides** [6] : Des architectures combinant CNN et RNN [14, 25] sont souvent utilisées pour bénéficier des avantages des deux types de réseaux : les CNN pour l'extraction de caractéristiques et les RNN pour la modélisation des séquences temporelles. Ces modèles convertissent directement un flux audio en une séquence de notes musicales, capturant à la fois les informations temporelles et les caractéristiques spectrales du signal.

4.4 Méthodes de Pointe : L'Approche de Google Brain

La méthode "state of the art" pour la transcription automatique, développée par l'équipe Google Brain [24], est principalement dédiée à la transcription du piano. Elle utilise une approche à double objectif en séparant le note tracking en deux parties : la détection du début des notes (onset detection) et la détection de la durée des notes (frame detection).

Cette méthode utilise une combinaison de réseaux de neurones convolutifs (CNN) et de variantes de réseaux de neurones récurrents (LSTM). Cette combinaison permet d'améliorer la précision de la transcription en se concentrant d'abord sur une détection d'onsets précise avec un premier réseau de neurones, puis en utilisant ces sorties (Onset Predictions), pour alimenter un second réseau chargé de détecter la durée des notes (Frame Predictions) (cf. Fig 4.2). Séparer le note tracking en deux étapes a permis d'adapter l'apprentissage aux différentes dynamiques temporelles des notes, ce qui a conduit à des résultats très précis.

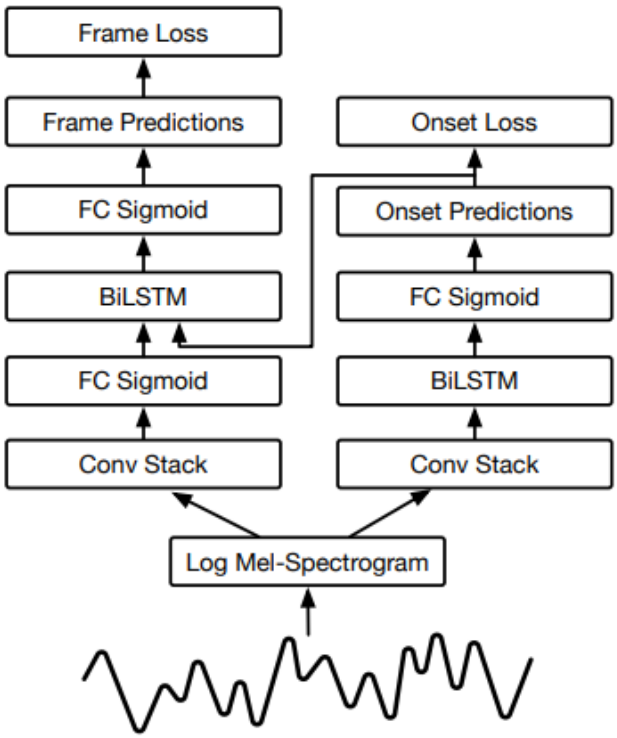

Figure 4.2: Architecture du modèle de transcription de Google Brain [24]

Chapter 5

Présentation de la méthode de transcription

Étant donné que l'objectif du stage était de proposer une méthode intégrant de l'intelligence artificielle, tout en tenant compte du fait que nous disposions de très peu de données, notre choix s'est orienté vers l'utilisation d'algorithmes et de modèles de machine learning déjà entraînés pour l'AMT d'instruments spécifiques, comme le piano, ainsi que pour la musique polyphonique.

Deux implémentations ont été choisies pour servir de méthodes de base (baselines) à partir desquelles nous évaluerons nos résultats. La première méthode, nommée Basic Pitch, est une bibliothèque Python dédiée à l'AMT. Elle utilise un réseau de neurones légers et efficaces, avec un nombre de paramètres réduit, ce qui la rend particulièrement adaptée aux environnements à faibles ressources.

5.1 Basic Pitch

Le modèle de Basic Pitch provient de l'article [35]. Cette méthode est conçue pour offrir une transcription musicale polyphonique et monophonique, tout en étant agnostique aux instruments, c'est-à-dire capable de transcrire des morceaux joués par différents instruments sans nécessiter de réentraînement [35]. La méthode repose sur un réseau de neurones convolutionnels qui prédit simultanément les activations des notes, les onsets (début des notes), et les multipitches à partir du signal audio.

La deuxième méthode est nommée Crepe Notes [34]. Elle propose une méthode innovante de transcription qui consiste à segmenter les contours de pitch détectés par un modèle existant appelé CREPE en notes discrètes. Nous allons détailler le processus de cette méthode.

5.2 CREPE notes

Le modèle CREPE Notes repose sur le modèle CREPE présenté dans l'article [9]. Il s'agit d'un système d'estimation de pitch de pointe utilisant un réseau de neurones convolutionnels (CNN) (cf. Figure 5.1). Ce modèle traite directement les échantillons audio bruts et prédit la fréquence fondamentale (f0) pour chaque tranche de 10 ms du fichier audio [9]. CREPE se distingue par sa robustesse et sa précision, même en présence de bruit,

surpassant ainsi des méthodes d'estimation de pitch reconnues telles que SWIPE [32] et PYIN [28], [9].

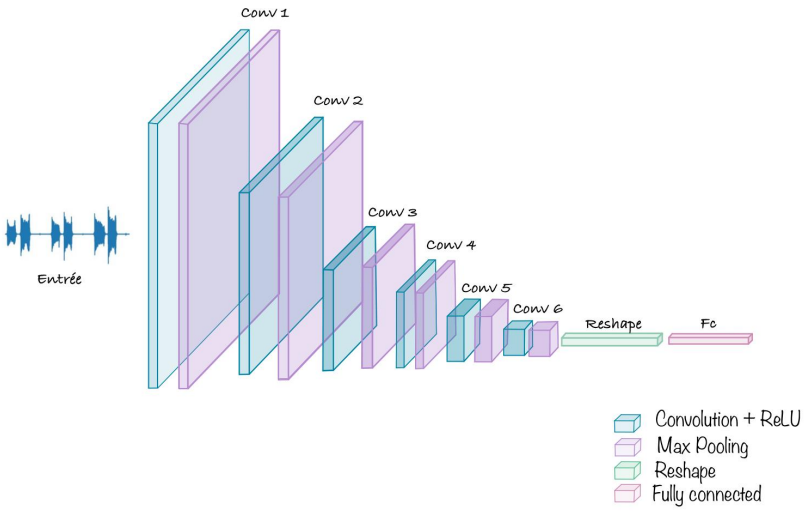

Figure 5.1: Architecture du modèle CREPE [9]

L'une des innovations majeures de CREPE Notes est l'utilisation combinée du gradient du pitch et de la mesure de confiance pour segmenter les notes musicales de manière précise [34]. La méthode est résumée sur le schéma ci-dessous (cf. Figure 5.2).

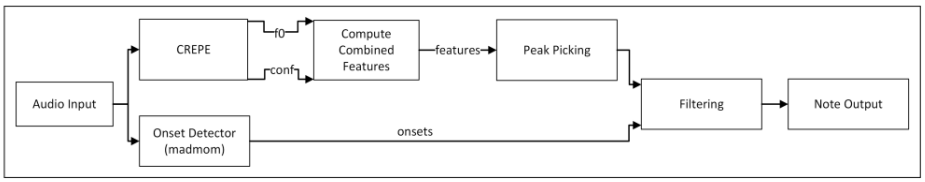

Figure 1. Flow diagram describing the main stages of the proposed method.

Figure 5.2: Méthode de segmentation CREPE Notes, image de l'article [34]

Le gradient du pitch mesure les variations de hauteur au fil du temps. Ce gradient est essentiel pour détecter les transitions entre les notes, c'est-à-dire les moments où le pitch change brusquement, correspondant souvent au début ou à la fin d'une note. La mesure de confiance fournie par CREPE indique à quel point le modèle est sûr de la prédiction du pitch à un moment donné. Elle est cruciale pour identifier les parties du signal où le pitch est bien défini. Cependant, cette mesure de confiance peut être bruitée, et elle est donc parfois insuffisante pour segmenter les notes de manière fiable lorsqu'elle est utilisée seule [34].

Pour surmonter les limitations de ces deux éléments lorsqu'ils sont utilisés séparément, CREPE Notes les combine. En multipliant le gradient du pitch normalisé par la mesure de confiance inversée, on obtient un signal plus clair (cf. Figure **??**). Ce signal combiné est ensuite utilisé pour identifier les frontières des notes, c'est-à-dire les moments où une note commence et se termine. Les pics dans ce signal, détectés avec une fonction de recherche de pics, correspondent généralement aux transitions entre les notes.

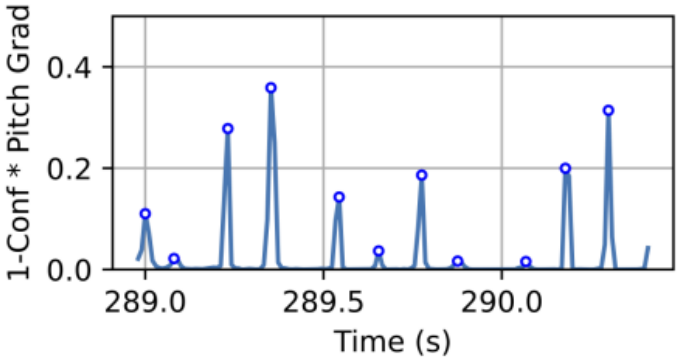

Figure 5.3: Signal combiné pour la segmentation des notes [34]

Un post-traitement est nécessaire pour gérer les notes répétées, trop courtes ou ayant une amplitude insuffisante. Pour les notes répétées, où le gradient de pitch reste constant, CREPE Notes fait appel à un modèle de détection d'onsets, le CNNOnsetProcessor de la librairie madmom. Ce réseau de neurones convolutionnel (CNN), décrit dans les articles [8] et [40], et dont le schéma est présenté dans la figure 5.4, permet de re-segmenter efficacement les notes en identifiant les onsets des notes, même lorsque les variations de pitch sont minimes ou inexistantes avec les prédictions de CREPE.

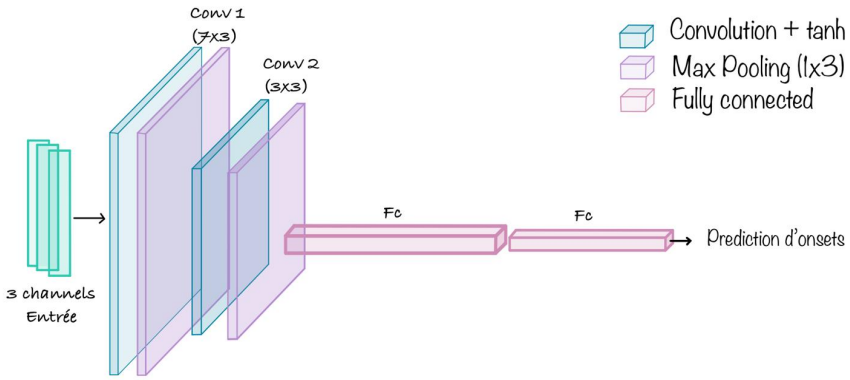

Figure 5.4: Architecture du modèle de détection d'onsets de madmom [8], [40]

Enfin, pour les notes courtes et faibles en amplitude, la méthode inclut également un seuil d'amplitude et de durée. Les débuts et fins des notes sont ajustés pour correspondre au profil d'amplitude perçu, garantissant ainsi une segmentation précise et fiable des notes musicales.

5.3 Amélioration de CREPE Notes

Dans le cadre de notre projet, nous avons apporté plusieurs modifications à la méthode CREPE Notes afin d'améliorer la précision de la détection des notes jouées sur la cithare *marovany*. L'une des principales modifications a consisté en l'ajout d'un filtrage plus strict sur l'amplitude, la durée, et le pitch des notes détectées. Sachant que nous connaissons

déjà à l'avance l'accordage spécifique des cordes de la cithare, nous sommes en mesure de restreindre la détection aux seules notes d'intérêt. Cette spécificité nous permet de tirer pleinement parti des informations disponibles et de concentrer la détection sur les onsets et offsets.

L'un des avantages de l'utilisation de CREPE Notes pour notre problème réside dans son utilisation du contour des pitches, une information directement liée aux fréquences des notes que nous connaissons déjà à l'avance. La méthode combine le gradient du pitch et la mesure de confiance pour former le **signal combiné** et affiner la détection des transitions entre les notes (cf. section 5.2). Bien que cela permette de détecter des transitions entre les notes, elle se contentait initialement d'utiliser un seul ensemble de pics (*peaks*) détectés à partir de ce signal combiné pour identifier les transitions (début et fin des notes). Nous avons ajouté une étape supplémentaire qui va plus loin en calculant les transitions non seulement à partir des pics du signal combiné (*peaks*), mais aussi à partir des pics de confiance (*conf_peaks*). Cette approche affine davantage les zones de transition, ce qui conduit à une détection plus précise des débuts et des fins de notes et limite les fausses détections. Cela est particulièrement utile dans les cas où la confiance est élevée mais le changement de pitch est subtil, ou vice versa. Par exemple, dans des situations où une note est soutenue mais où son pitch change peu, la confiance du modèle peut indiquer une transition significative qui serait autrement ignorée par le simple gradient de pitch.

5.4 Ré-entrainement du CNN de Madmom

Après avoir observé les premiers résultats de transcription avec CREPE Notes sur nos données de cithare, nous avons constaté que de nombreux onsets n'étaient pas détectés. Ce problème a été attribué en partie à l'utilisation du modèle CNN de Madmom pour la détection d'onsets, qui, bien qu'il soit très efficace, est un peu ancien (2016). Nous avons donc décidé de le remplacer par un CNN un peu plus récent, entraîné spécifiquement sur notre base de données, qui serait plus adapté à la détection des différents types de notes jouées sur la *marovany*, qui peuvent varier considérablement en termes de durée et d'amplitude (cf. figure 3.2).

Pour ce faire, nous avons repris l'architecture de base du CNN de Madmom, en y ajoutant plusieurs améliorations :

- **Batch Normalization** : Après chaque couche de convolution, nous avons ajouté des couches de Batch Normalization pour stabiliser et accélérer l'entraînement. Cela réduit la sensibilité du réseau aux variations d'initialisation et diminue les risques de surapprentissage.
- **Augmentation du nombre de filtres** : Nous avons augmenté le nombre de filtres dans les couches convolutionnelles (32 filtres dans la première couche et 64 dans la deuxième) pour permettre au modèle de capturer des caractéristiques plus complexes.
- **Fonction d'Activation ReLU** : Nous avons opté pour la fonction ReLU, qui est plus populaire dans les architectures modernes en raison de sa convergence plus rapide et de sa meilleure gestion des gradients.

- **Régularisation L2** : Nous avons ajouté une régularisation L2 pour pénaliser les poids excessifs et ainsi éviter le surapprentissage. Cette technique aide à simplifier le modèle tout en conservant sa capacité à apprendre des représentations riches.

5.4.1 Comparaison des Architectures

L'architecture reste globalement la même (cf. figure 5.4) : elle prend en entrée trois mel-spectrogrammes, a le même nombre de couches de convolution et de couches entièrement connectées. Cependant, avec ses améliorations, elle est plus complexe et capable de capturer des caractéristiques plus détaillées, ce qui la rend mieux adaptée à la détection d'onsets dans notre contexte spécifique du jeu de la cithare, qui est assez complexe. Les modifications apportées, telles que l'ajout de la Batch Normalization et l'augmentation du nombre de filtres, permettent de stabiliser l'entraînement et d'améliorer la performance du modèle. Le choix de la fonction ReLU renforce également l'efficacité de l'apprentissage, en particulier pour des réseaux profonds comme le nôtre. L'architecture du nouveau CNN est présenté figure 5.5

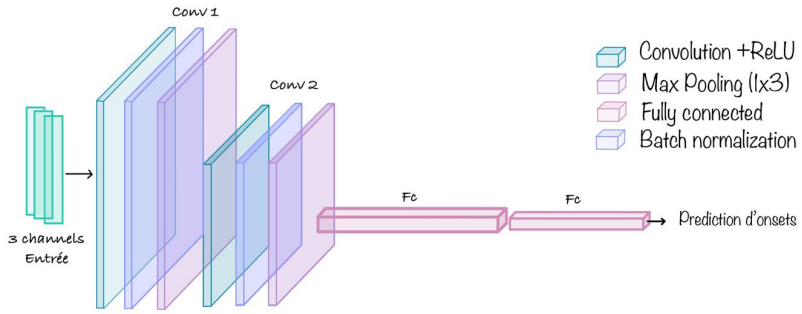

Figure 5.5: Architecture du modèle de détection d'onsets de madmom [8], [40]

Chapter 6

Expérimentation

6.1 La Dataset

Le grand défi de ce projet réside dans le nombre très réduit de données disponibles. Nous disposons de deux jeux de données enregistrés sur la cithare *marovany* :

- *Velonjoro*, qui comprend 11 cordes enregistrées sur 5 minutes 14 secondes.
- *Sojerina*, qui comprend 13 cordes enregistrées sur 5 minutes 10 secondes.

Étant donné que nous avons enregistré chaque corde séparément grâce à des capteurs piézoélectriques placés sur chacune d'elles, nous pouvons considérer chaque corde comme un enregistrement monophonique. Ainsi, chaque corde est traitée comme une musique indépendante.

En tout, cela nous donne 24 enregistrements, soit un peu plus de deux heures d'enregistrement au total, ce qui reste relativement limité pour un projet de cette envergure. Afin d'élargir cette petite base de données, nous avons mis en place différentes méthodes d'augmentation des données.

6.1.1 Segmentation et Annotation Semi-Automatique

Pour faciliter l'annotation manuelle, nous avons commencé par diviser chaque enregistrement en segments de 10 secondes.

L'interface graphique, utilisée lors du stage de M1 (cf. Introduction), a été légèrement modifiée afin de permettre une annotation semi-automatique. Cette application, qui reposait sur des méthodes de traitement du signal pour détecter les notes, incluait également des outils pour modifier manuellement les résultats [46]. Concrètement, nous générions une détection automatique avec cet algorithme, puis corrigions manuellement les erreurs. Bien que cette méthode ne suffise pas pour une transcription automatique complète, elle facilite considérablement le travail d'annotation en détectant une partie des notes. Exemple est illustrer figures 6.2 et 6.2.

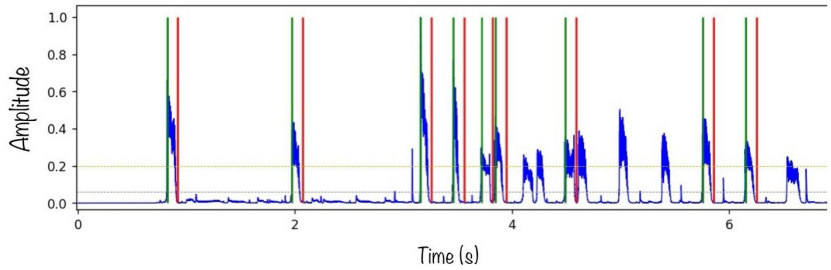

Figure 6.1: Détection de notes effectuée par l'algorithme.

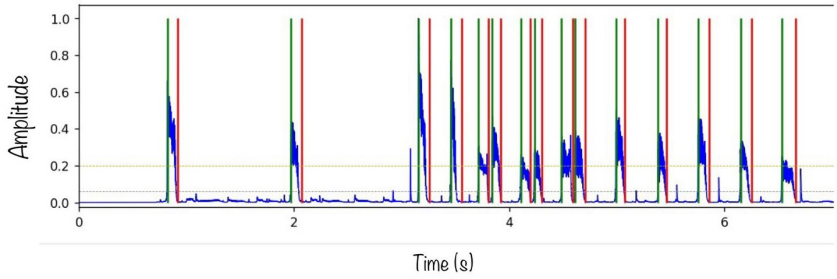

Figure 6.2: Détections après correction manuelle.

6.1.2 Génération de Données Synthétiques

Pour augmenter la base de données, nous avons également créé de nouveaux enregistrements de manière synthétique. Bien entendu, ces enregistrements ne respectent pas les caractéristiques musicales du jeu du cithariste, notamment son placement rythmique par rapport à une pulsation. Mais dans la pratique des réseaux de neurones en transcription automatique, l'utilisation de données synthétiques permet d'améliorer les performances du système même si elle ne sont pas strictement conformes au style musical étudié.

Nous avons implémenté un script Python qui, dans un premier temps, isole chaque note des enregistrements corde par corde grâce aux annotations. Ensuite, il crée de nouveaux fichiers audio (`.wav`) en mixant ces notes de manière aléatoire. Afin de respecter autant que possible le style de jeu des citharistes, nous avons varié les intervalles entre les notes, allant de très courts à plus longs. En effet, dans le jeu de la cithare, il est courant que plusieurs notes soient jouées très rapidement les unes après les autres, suivies parfois de silences plus longs.

6.1.3 Ajout de Bruits pour Diversification

Pour élargir encore davantage la base de données et permettre à notre modèle de mieux généraliser, nous avons utilisé deux méthodes d'ajout de bruit aux mix créés :

- **Ajout de bruit blanc** : Nous avons ajouté du bruit blanc de manière aléatoire à certains fichiers de la dataset, avec une probabilité de 30% que du bruit soit présent dans un mix. Plusieurs niveaux de rapport signal/bruit (SNR) ont été utilisés : 5, 10, 15, 20 et 30 dB. On peut voir exemple de mix bruité figure 6.3
- **Ajout de bruit extrait des enregistrements réels** : Nous avons également extrait le bruit des enregistrements originaux en nous basant sur les annotations,

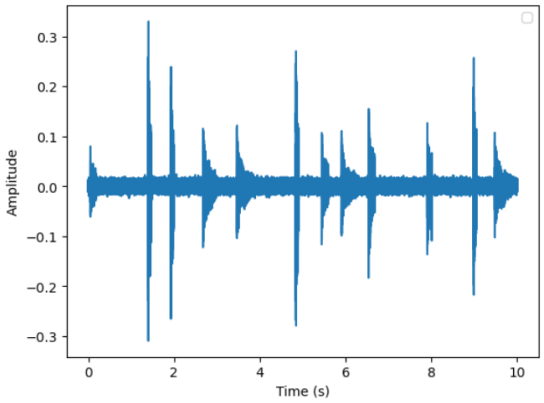

Figure 6.3: Exemple de mix synthétique généré.

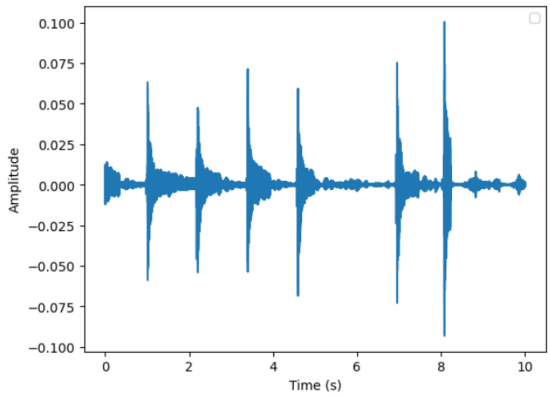

Figure 6.4: Séquence enregistrée réelle.

en supposant que les segments sans notes contenaient principalement du bruit ou des résonances. Ce bruit a ensuite été ajouté de manière aléatoire dans la dataset synthétique.

Ces deux approches permettent d'augmenter la base de données tout en la diversifiant: d'une part, avec un bruit théorique (le bruit blanc), et d'autre part, avec un bruit plus réaliste, issu de l'électronique, des capteurs ou encore de l'environnement (bruits de la salle ou des personnes présentes).

6.2 Pré-traitements des données

Le prétraitement des données reste le même que celui décrit dans l'article [8] : trois Mel-spectrogrammes sont générés pour chaque fichier audio en utilisant différentes tailles de fenêtre FFT (1024, 2048, 4096), permettant ainsi de capturer des informations à diverses résolutions temporelles et fréquentielles.

Un Mel-spectrogramme est une représentation visuelle du spectre des fréquences d'un signal audio, transformée en échelle Mel pour mieux correspondre à la perception humaine des fréquences. Contrairement à un spectrogramme classique, où l'échelle des fréquences est linéaire, l'échelle Mel est logarithmique, ce qui accorde plus de détails aux basses fréquences et réduit l'importance des hautes fréquences, conformément à la façon dont l'oreille humaine perçoit les sons.

L'utilisation de différentes tailles de fenêtre FFT (1024, 2048, 4096) est pertinente pour capturer à la fois les caractéristiques temporelles et fréquentielles du signal audio :
FFT de 1024 : Cette petite taille de fenêtre offre une bonne résolution temporelle, ce qui permet de capturer avec précision les événements sonores rapides, tels que les attaques de notes. FFT de 2048 : Elle représente un compromis entre la résolution temporelle et fréquentielle, offrant une vue équilibrée du signal. FFT de 4096 : Avec cette grande fenêtre, la résolution fréquentielle est améliorée, permettant de distinguer les harmoniques proches les unes des autres, mais au détriment de la résolution temporelle.

Ces spectrogrammes sont ensuite convertis en échelle logarithmique pour aligner la représentation des données avec la perception auditive humaine. La normalisation des

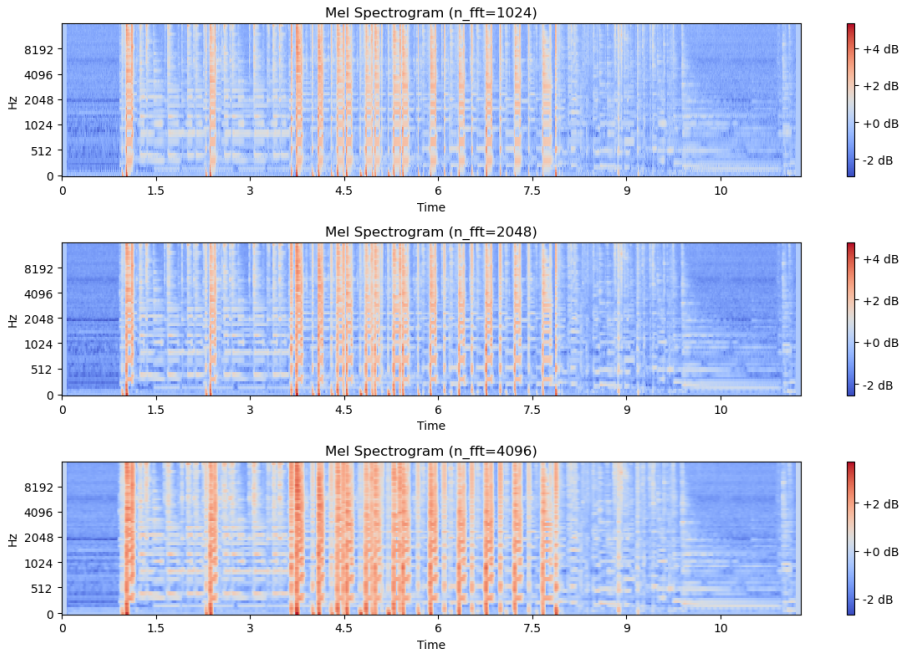

Figure 6.5: Exemple de mel-spectrogrammes

spectrogrammes garantit que les données sont centrées et réduites, facilitant ainsi l'apprentissage par les réseaux de neurones. Enfin, les spectrogrammes sont découpés en segments temporels fixes avec chevauchement, fournissant un contexte temporel pour chaque décision d'onset. Sur la figure 6.5, on peut voir un exemple de mel-spectrogrammes qu'on a en entrée du réseau.

6.3 Entraînement

Concernant l'entraînement, nous avons conservé la même méthode que celle décrite dans l'article de référence [8].

Le modèle prédit un 1 lorsqu'un onset (début de note) est détecté et un 0 sinon. Cette détection correspond à une classification binaire où les classes sont soit 0, soit 1. Dans un enregistrement de cithare, il y a davantage d'instants sans onset que d'instants où un onset est présent. Par conséquent, la base de données contient plus de labels 0 que de labels 1. Cela crée un déséquilibre des classes : la classe majoritaire est celle des 0, et la classe minoritaire celle des 1. Ce déséquilibre risque d'impacter négativement le modèle, qui pourrait ainsi apprendre à prédire principalement l'absence de notes.

Pour remédier à ce problème, nous avons choisi d'ajouter une fonction de *under-sampling*, qui consiste à sous-échantillonner la classe majoritaire (la classe 0) pour rééquilibrer les classes et obtenir un nombre de 0 presque équivalent à celui des 1 (la classe minoritaire). Cela permet d'éviter que le modèle soit biaisé en faveur de la classe dominante. Cette méthode d'équilibrage est particulièrement utile pour des données déséquilibrées comme les nôtres, où les événements (onsets) sont souvent rares par rapport aux segments sans onset détecté.

Nous avons également ajouté un mécanisme d'*early stopping* afin de surveiller la performance du modèle sur l'ensemble de validation et d'arrêter l'entraînement si aucune amélioration significative n'est observée. Cela permet de prévenir le surapprentissage, que nous avions remarqué lors des premiers essais, en s'assurant que le modèle ne continue pas à s'entraîner inutilement, ce qui pourrait entraîner une dégradation des performances sur de nouvelles données.

Chapter 7

Résultats

7.1 Évaluation

Pour évaluer la performance de notre modèle de détection des onsets et des notes, nous avons adopté l'approche décrite dans l'article [8] basée sur la validation croisée en huit plis (*8-fold cross-validation*). Cette méthode consiste à diviser l'ensemble des données en huit sous-ensembles égaux. À chaque itération, sept sous-ensembles sont utilisés pour l'entraînement du modèle, tandis que le sous-ensemble restant est utilisé pour l'évaluation. Ce processus est répété huit fois, chaque sous-ensemble étant utilisé une fois pour l'évaluation et sept fois pour l'entraînement. Cette approche permet d'obtenir une estimation plus fiable des performances du modèle sur des données non vues, en maximisant l'utilisation des données disponibles pour l'entraînement tout en fournissant une évaluation robuste.

Les métriques finales sont ensuite obtenues en moyennant les résultats des huit plis. Nous utilisons `mir_eval` pour calculer plusieurs métriques clés, notamment la précision, le rappel, le score F1 et le taux de chevauchement (*overlap ratio*). Ces métriques sont calculées pour la détection des onsets seule ainsi que pour la transcription globale (onsets + offsets + pitch). Elles sont définies en fonction de la correspondance entre les notes prédites et les notes de référence, avec une tolérance pour les décalages d'onset et d'offset afin de tenir compte des petites variations temporelles. Les définitions des principales métriques sont les suivantes :

- **Précision (Precision)** : Proportion des onsets (ou des notes) prédits qui sont corrects.
- **Rappel (Recall)** : Proportion des onsets (ou des notes) de référence qui sont correctement détectés par le modèle.
- **Score F1 (F1 Score)** : Moyenne harmonique de la précision et du rappel. Elle fournit une seule valeur qui prend en compte à la fois la précision et le rappel.
- **Taux de Chevauchement (Overlap Ratio)** : Mesure de la similitude temporelle entre les notes correspondantes.

7.2 Résultats

Les tableaux 7.1 et 7.2 présentent respectivement les résultats de transcription globale (onset + offset + pitch) et les résultats spécifiques aux détections d'onsets pour différentes

Table 7.1: Résultats de transcription

Méthode	**Précision (%)**	**Rappel (%)**	**F1 Score (%)**	**Overlap Ratio (%)**
Baseline	62.14	**85.46**	71.96	**77.33**
CN Modifié	68.61	**84.93**	**75.90**	**76.38**
CN Modifié + Post	**75.28**	75.89	**75.58**	**76.94**
NewCNN	56.69	46.38	51.02	75.65
NewCNN + Post	62.45	41.69	50.00	75.91

Table 7.2: Résultats détections d'onsets

Méthode	**Précision (%)**	**Rappel (%)**	**F1 Score (%)**
Baseline	71.36	**98.13**	82.63
CN Modifié	77.05	95.38	85.24
CN Modifié + Post	86.31	87.01	**86.66**
NewCNN	86.84	71.06	78.16
NewCNN + Post	**93.71**	62.55	75.02

méthodes : Baseline, CN Modifié (*CREPE Notes Modifié*, cf. partie 5.3), CN Modifié avec Post-Traitement, NewCNN (nouveau CNN, cf. partie 5.4)et NewCNN avec Post-Traitement. Ces méthodes ont été évaluées sur des enregistrements réels de jeux de cithare.

Les résultats de la Baseline *CREPE Notes* montrent une performance globalement satisfaisante, avec un F1 Score de 71.96%. Cependant, cette méthode présente un déséquilibre notable entre la précision et le rappel, avec un rappel très élevé mais une précision modérée. Ce déséquilibre peut entraîner de nombreux faux positifs, où des onsets sont détectés là où il n'y en a pas.

1. Perfomances de NewCNN

Le *NewCNN* a été conçu pour améliorer la détection des onsets dans la musique de cithare et, par conséquent, améliorer les transcriptions en utilisant une architecture CNN moderne. Cependant, malgré ces efforts, les résultats montrent que la méthode *NewCNN* sans post-traitement n'a pas réussi à surpasser la baseline. Par exemple, le F1 Score est de 51.02% contre 71.96% pour la baseline. Bien que le modèle offre une meilleure précision dans certains cas, notamment pour la détection des onsets où il obtient une précision de 93.71%, il souffre d'une baisse significative du rappel, ce qui indique qu'il manque de nombreux onsets.

Avec l'ajout du post-traitement pour supprimer les fausses détections (faux positifs) (méthode *NewCNN + Post-Traitement*), la performance s'améliore légèrement, mais reste inférieure à celle de la baseline. Cela suggère que, bien que le *NewCNN* soit capable d'apprendre des caractéristiques complexes, il nécessite encore des ajustements.

L'une des raisons pour lesquelles les résultats ne dépassent pas la baseline est liée à la base de données. Les résultats d'entraînement du modèle montrent qu'il généralise bien sur la base de données synthétique, mais ce n'est pas le cas avec les jeux de données de cithare. En effet, la base de données étant synthétique (cf. partie 6.1), elle n'est pas

suffisamment représentative des vrais jeux de cithare malgré tous les efforts mis en place. Par conséquent, lorsque le modèle rencontre de vraies données, il a plus de difficulté à détecter les onsets.

2. Résultats de CREPE Notes Modifié

La méthode *CREPE Notes Modifié* (CN Modifié) a été introduite pour exploiter les pics de confiance (*conf_peaks*) et un filtrage plus strict en plus des informations traditionnelles utilisées par la baseline (cf. parties 5.2 et 5.3). Cette méthode a montré des améliorations notables par rapport à la baseline, avec un F1 Score atteignant 75.90%, surpassant largement celui de la baseline. Cela démontre que l'intégration des pics de confiance dans le processus de détection des onsets améliore la précision de la transcription.

De plus, avec l'ajout d'un post-traitement spécifique qui permet de supprimer les faux positifs (*CN Modifié + Post-Traitement*), la performance se stabilise, avec un F1 Score de 75.58% et un *Onset F-measure* de 86.66%. Cette approche combinée offre une détection des onsets plus robuste et plus fiable, surpassant les performances de la baseline de manière significative.

Pour finir, les méthodes *NewCNN*, *CN Modifié* et *CN Modifié + Post-Traitement* surpassent largement l'application développée lors du stage de M1, qui nécessitait beaucoup trop de modifications manuelles. Ces nouvelles méthodes atteignent de plus hautes performances sans nécessiter de modifications manuelles.

3. Résultats de Transcription pour les Enregistrements des Micros Généraux

Les résultats des transcriptions basées sur les enregistrements des micros généraux, qui capturent toutes les cordes de la cithare simultanément, montrent des performances globalement inférieures par rapport aux enregistrements séparés des cordes. En moyenne, le *F1 Score* est de 34.18%. Ces résultats soulignent la difficulté accrue de la tâche de transcription lorsqu'elle est appliquée à des enregistrements polyphoniques, où plusieurs cordes sont capturées simultanément, particulièrement pour la cithare *marovany*, où les notes sont généralement jouées très rapidement.

L'utilisation de micros généraux transforme la tâche de transcription en un problème polyphonique complexe, où le modèle doit identifier et distinguer plusieurs notes jouées simultanément. Cela explique les performances médiocres observées dans les résultats ci-dessus. En revanche, les enregistrements séparés, qui rendent chaque piste monophonique, simplifient considérablement la tâche de transcription (cf. Tab.7.1 et 7.2), permettant au modèle de se concentrer sur une seule note à la fois. D'où l'importance ici d'avoir des enregistrements séparés pour chaque corde de la cithare.

Chapter 8

Conclusion

Les modifications apportées au modèle CREPE Notes, notamment les méthodes CN Modifié et CN Modifié avec Post-Traitement, ont permis d'obtenir de meilleurs résultats de transcription sur la base de données de la cithare. Ces améliorations se sont avérées efficaces car elles intègrent les pics de confiance, des filtrages plus stricts et un post-traitement dans le processus de détection des notes, améliorant ainsi la précision des transcriptions et réduisant les faux positifs. En revanche, la modification du CNN pour la détection des onsets (NewCNN) n'a pas permis de surpasser les résultats de la baseline. Cependant, les performances du NewCNN dépassent néanmoins celles obtenues avec la première approche développée lors de mon stage de M1. Cette amélioration est significative, bien qu'elle démontre également que des ajustements supplémentaires sont nécessaires pour affiner ce modèle.

Les perspectives de ce projet sont multiples, mais l'une des améliorations principales à apporter concerne la base de données. Le manque de données annotées pour la transcription reste aujourd'hui une barrière majeure au développement de modèles d'intelligence artificielle. Cela est particulièrement vrai pour la cithare *marovany*, qui ne possède aucune base de données officielle. La prochaine étape serait donc de créer un dataset à partir de véritables enregistrements de cithare, offrant une diversité accrue qui permettrait aux réseaux de neurones de mieux généraliser.

Avec une base de données plus riche, il serait alors envisageable de s'orienter vers des modèles plus complexes, tels que celui implémenté par l'équipe de Google Brain, qui combinent un CNN et un LSTM (cf. article [24]). Ce modèle représente l'état de l'art actuel de la transcription automatique et pourrait potentiellement offrir des performances supérieures.

Bilan Personnel

Les difficultés rencontrées lors de ce projet ont été nombreuses. La revue de l'état de l'art a été longue, et décider des méthodes les plus appropriées pour notre cas spécifique n'a pas été simple. La création de la dataset a également été un défi majeur : les annotations ont été longues et fastidieuses, et la création de données synthétiques a nécessité de nombreuses études pour tenter de se rapprocher au maximum d'un vrai jeu de cithare.

Malgré ces difficultés, ce stage m'a permis d'acquérir de nombreuses compétences que

je ne me pensais pas capable de développer auparavant. C'était la première fois que je réalisais un projet d'une telle envergure en autonomie, ce qui a rendu les prises de décision particulièrement difficiles car je manquais de confiance en mes idées. Cependant, cette expérience m'a appris à avoir confiance en mes décisions tout en étant capable de les remettre en question. J'ai également appris à évaluer la faisabilité d'une méthode, une compétence qui s'est avérée cruciale pendant ce stage. En consacrant du temps à l'étude de l'état de l'art et des architectures existantes, j'ai pu mieux juger de la possibilité d'implémenter certaines méthodes dans un délai limité.

En conclusion, je pense que ce projet m'a permis d'évoluer considérablement sur les plans personnel, professionnel et technique. Cette expérience a été enrichissante et m'a apporté de nombreuses connaissances, qui me seront très utiles à l'avenir.

Bibliography

[1] A. Klapuri and M. Davy, Eds., *Signal Processing Methods for Music Transcription.* New York: Springer, 2006.

[2] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, 2019.

[3] S. Sigtia, E. Benetos, and S. Dixon, "An end-to-end neural network for polyphonic piano music transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 927–939, May 2016.

[4] S. Böck and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 121–124.

[5] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription," in *Proc. International Conference on Machine Learning (ICML)*, 2012.

[6] S. Sigtia, E. Benetos, N. Boulanger-Lewandowski, T. Weyde, A. S. d'Avila Garcez, and S. Dixon, "A Hybrid Recurrent Neural Network for Music Transcription," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, Australia, April 2015, pp. 2061–2065.

[7] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, and G. Widmer, "On the potential of simple framewise approaches to piano transcription," in *Proc. International Society for Music Information Retrieval Conference*, 2016, pp. 475–481.

[8] J. Schlüter and S. Böck, "Improved musical onset detection with convolutional neural networks," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 6979–6983.

[9] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 161–165.

[10] K. Ullrich, J. Schlüter, and T. Grill, "Music transcription with convolutional sequence-to-sequence models," in *Proc. 18th International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou, China, 2017, pp. 23–29.

[11] S. Böck and M. Schedl, "Polyphonic piano note transcription with recurrent neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, 2012, pp. 121–124.

[12] J. Gardner, I. Simon, E. Manilow, C. Hawthorne, and J. Engel, "MT3: Multi-task multitrack music transcription," in *Proc. International Conference on Learning Representations (ICLR)*, Virtual Event, 2022.

[13] C. Hawthorne, I. Simon, R. Swavely, E. Manilow, and J. Engel, "Sequence-to-sequence piano transcription with transformers," in *Proc. 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[14] M. Mínguez Carretero, "Automatic music transcription using neural networks," M.S. thesis, Dept. Computer Science, University of Alicante, Spain, 2018.

[15] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.

[16] L. Su and Y.-H. Yang, "Combining spectral and temporal representations for multipitch estimation of polyphonic music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1600–1612, Oct 2015.

[17] P. H. Peeling, A. T. Cemgil, and S. J. Godsill, "Generative spectrogram factorization models for polyphonic piano transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 519–527, March 2010.

[18] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[19] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 528–537, 2010.

[20] E. Benetos and S. Dixon, "Multiple-instrument polyphonic music transcription using a temporally-constrained shift-invariant model," *Journal of the Acoustical Society of America*, vol. 133, no. 3, pp. 1727–1741, March 2013.

[21] B. Fuentes, R. Badeau, and G. Richard, "Harmonic adaptive latent component analysis of audio and application to music transcription," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1854–1866, Sept 2013.

[22] J. Nam, J. Ngiam, H. Lee, and M. Slaney, "A classification-based polyphonic piano transcription approach using learned feature representations," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2011, pp. 175–180.

[23] M. Marolt, "A connectionist approach to automatic transcription of polyphonic piano music," *IEEE Transactions on Multimedia*, vol. 6, no. 3, pp. 439–449, 2004.

[24] A. Cogliati, Z. Duan, and B. Wohlberg, "Context-dependent piano music transcription with convolutional sparse coding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2218–2230, Dec 2016.

[25] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," in *Proc. International Society for Music Information Retrieval Conference*, 2018.

[26] A. De Cheveigne and H. Kawahara, “Yin, a fundamental frequency estimator for speech and music,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[27] A. Ycart and E. Benetos, “Polyphonic music sequence transduction with meter-constrained LSTM networks,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 386–390.

[28] M. Mauch and S. Dixon, “pYIN: A fundamental frequency estimator using probabilistic threshold distributions,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 659–663.

[29] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, "madmom: A new Python audio and music signal processing library," in *Proc. 24th ACM International Conference on Multimedia*, Amsterdam, Netherlands, 2016, pp. 1174-1178.

[30] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, “Discriminative Non-Negative Matrix Factorization For Multiple Pitch Estimation,” in *Proc. International Conference on Music Information Retrieval (ISMIR)*, Porto, Portugal, 2012, pp. 205–210.

[31] A. Dessein, A. Cont, and G. Lemaitre, "Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence," in *Proc. 11th International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, Netherlands, 2010, pp. 489-494.

[32] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," in *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[33] Z. Meng and W. Chen, "Automatic music transcription based on convolutional neural network, constant Q transform and MFCC," in *Journal of Physics: Conference Series*, vol. 1651, no. 1, p. 012192, 2020.

[34] X. Riley and S. Dixon, "CREPE Notes: A new method for segmenting pitch contours into discrete notes," in *Proc. 22nd International Society for Music Information Retrieval Conference (ISMIR)*, Online, 2021.

[35] M. Bittner, J. J. Bosch, D. Rubinstein, G. Meseguer-Brocal, and S. Ewert, "A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022, pp. 781–785.

[36] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, “Automatic music transcription: challenges and future directions,” *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[37] P. Smaragdis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[38] S. B. Puri and S. Mahajan, “Review on automatic music transcription system,” in *Proc. International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 2017, pp. 1–6.

[39] N. Obin, “Advanced Audio Processing: C5. Audio Source Separation”,in *Master in Engineering of Intelligent Systems*, Sorbonne Université, 2021-2022.

[40] J. Schlüter and S. Böck, “Musical onset detection with convolutional neural networks,” in *6th International Workshop on Machine Learning and Music (MML)*, Prague, Czech Republic, 2013.

[41] M. Mauch and S. Dixon, “PYIN: A fundamental frequency estimator using probabilistic threshold distributions,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 659–663. [Online]. Available: `http://ieeexplore.ieee.org/document/6853678/`

[42] D. Cazau, “Automatic Music Transcription based on Prior Knowledge from Musical Acoustics. Application to the repertoires of the marovany zither of Madagascar,” Ph.D. dissertation, Discipline: Acoustics and Signal Processing, 2015.

[43] D. Cazau, Y. Wang, M. Chemillier, and O. Adam, “An automatic music transcription system dedicated to the repertoires of the marovany zither,” *Journal of New Music Research*, vol. 45, no. 4, pp. 343–360, 2016.

[44] M. Chemillier and J. Vali, “Bi-musicality in the Age of Artificial Intelligence,” in *SoMoS, ICTM Study Group on Sound, Movement, and the Sciences*, Barcelona, October 26, 2022, pp. 5-9. [Online]. Available: `http://ehess.modelisationsavoirs.fr/marc/publi/Chemillier_SoMos-2022-CopyEdited_KVFBB-Chemillier-preprint.pdf`

[45] M. Chemillier, “The smile of Velonjoro. Bi-musicality and the use of artificial intelligence in the analysis of Madagascar zither music,” *Analytic Approach to African Music*, à paraître 2025. [Online]. Available: `http://ehess.modelisationsavoirs.fr/marc/publi/Analytic-Approach-African-Music-Chemillier-v2-preprint.pdf`

[46] L.Rahaoui, "Etude de paramétrage d’un algorithme de Transcription Automatique et de détection MIDI pour la cithare Marovany",Rapport de stage Master 1 Automatique Robotique, 2023