Louis Lumière
école nationale supérieure

La Cité du Cinéma – 20 Rue Ampère BP 12 – 93213 La Plaine Saint-Denis
Téléphone : +33 (0)1 84 67 00 01 – https://www.ens-louis-lumiere.fr

Mémoire de Master

Spécialité Son, Promotion 2022

CLASSIFICATION EN TEMPS RÉEL DE TECHNIQUES EXTRÊMES DE DISTORSION VOCALE UTILISÉES DANS LE HEAVY METAL

Modan Tailleur
modan.tailleur@gmail.com

Directeur interne et Président du Jury : Laurent Millot, Maître de Conférences en traitement du signal

Directeur externe : Julien Pinquier, Maître de Conférences

Coordinateur des Mémoires : Corsin Vogel, PAST

Rapporteur : Frank Gillardeaux

Résumé Les techniques de saturation vocale sont très communément employées par les chanteurs de musique *metal*, particulièrement dans les sous-genres les plus extrêmes. De multiples techniques peuvent être utilisées au sein d'un même morceau par les chanteurs. Ces techniques peuvent bénéficier de traitements radicalement différents de la part des ingénieurs du son, traitements qui sont particulièrement difficiles à mettre en œuvre lors de performances live. Ce mémoire constitue une étude préliminaire au développement d'un plug-in qui permettrait de détecter en temps réel différentes techniques extrêmes de saturation vocale, et de rediriger le signal vers des bus de traitement adaptés à chaque technique. Après avoir présenté une nouvelle taxonomie des chants extrêmes saturés, plusieurs méthodes de *Machine Learning* sont explorées à partir d'enregistrements de voix de 27 chanteurs et chanteuses de *metal* : le perceptron multicouche, la forêt d'arbres décisionnels, et la classification naïve bayésienne. Lors de ce travail, de nouveaux descripteurs acoustiques nommés DAFCC (Data Adjusted Frequencies Cepstral Coefficients) ont été élaborés afin de s'adapter au mieux aux données du problème. Ces descripteurs sont directement inspirés des MFCC. L'extraction des DAFCC, comparée à celle des MFCC, permet de passer d'une précision de 74,5% à une précision de 75,5% tout en réduisant le temps de calcul de l'algorithme. En créant des modèles personnalisés pour chaque sous-genre du *heavy metal*, la précision atteint un score variant entre 77,9% et 96,4%. Ces scores de précision, obtenus à partir du perceptron multicouche exploitant des trames de 1024 échantillons, montrent beaucoup de potentiel pour le développement futur d'un programme susceptible de fonctionner en temps-réel.

Mots Clés : voix chantée, distorsion vocale, rugosité vocale, DAFCC, MLP, temps-réel.

Abstract Vocal distortion techniques are very commonly used by metal singers, especially in the more extreme sub-genres. Singers can sometimes use multiple techniques within the same piece of music. These techniques can be processed very differently by the sound engineers, and using different sound effects on each technique can be particularly difficult to implement during live performances. This master's thesis is a preliminary study for the development of a plug-in that would be able to detect in real time different extreme vocal distortion techniques, and to redirect the signal to processing buses adapted to each technique. After presenting a new taxonomy of extreme distorted vocals, several Machine Learning methods are explored based on voice recordings of 27 metal singers : the multilayer perceptron, the random forest, and the Gaussian Naive Bayes classification. In this work, new acoustic features named DAFCC (Data Adjusted Frequencies Cepstral Coefficients) were developed in order to best fit the data. These descriptors are directly inspired by the MFCC. The extraction of DAFCCs, compared to MFCCs, leads to an improvement of the accuracy while reducing the computation time of the algorithm. The accuracy is 74.5% with the use of MFCCs, and 75.5% with the use of DAFCCs. By creating custom models for each heavy metal sub-genre, the accuracy scores range from 77.9% to 96.4%. These accuracy scores, obtained from the multilayer perceptron using 1024 sample frames, show a lot of potential for the future development of a program able to perform in real-time.

Key Words : singing voice, vocal distortion, vocal roughness, DAFCC, MLP, real-time.

Remerciements

Je tiens à remercier tous les participants ayant prêté leur voix au projet : Xavier Hedan, Larry Etienne, Joachim Preschner, Joshua Smith, Samuel Girard, Ingrid Fleury, Arnold Petit, Emil Azzam, Nicolas Foucault, Sébastien Tuvi, Romain Le Bihan, Julien Deyres, Sébastien Scherrer, Simon Perrin, Alexandre Lheritier, Simon Houdin, Ricardo Gomes, Laetitia Jehano, Didier Cauchois, Marine Ternon, Florian Gatta, Antonin Philippe, Matthieu Baudiment, Martin Rambaud, Yasmine Liverneaux, Antoine Dumortier et Philippe Charny.

Un grand merci tout particulièrement à Joshua Smith, qui m'a beaucoup aidé dans la rédaction d'une nouvelle taxonomie de voix saturées.

Merci à Christophe D'Alessandro, qui m'a aidé à cerner ce sujet de mémoire, et à Geoffroy Peeters qui m'a aiguillé au début de ce projet.

Merci à Oriol Nieto qui par son expérience m'a beaucoup aidé à appréhender ce travail.

Merci à Laurent Millot, Julien Pinquier, et Corsin Vogel pour leur suivi tout au long de ce mémoire.

Merci à Margaux Limbour pour ses multiples relectures.

Glossaire

black metal un des nombreux sous-genres extrêmes du *heavy metal*, ayant développé ses propres techniques vocales saturées (voir tableau 1.1). 11, 13, 14, 16, 41–43, 46, 81, 82

black shriek le *black shriek* (BS) est une catégorie de distorsion vocale utilisée dans ce document pour désigner la technique vocale majoritairement utilisée dans le *black metal*. Sorte de cri perçant (voir description en section 2.1). 41–43, 55, 57, 62, 64, 73–76, 79–82

death growl le *death growl* (DG) est une catégorie de distorsion vocale utilisée dans ce document pour désigner la technique vocale majoritairement utilisée dans le *death metal*. Extrêmement grave et saturé (voir description en section 2.1). 41–44, 55, 57, 63, 67, 73–75, 79, 80, 82

death metal un des nombreux sous-genres extrêmes du *heavy metal*, ayant développé ses propres techniques vocales saturées (voir tableau 1.1). 11, 13, 14, 16, 41, 42, 46, 80–82

deep gutturals le *deep gutturals* (DeG) est un *death growl* très grave utilisé comme effet (voir description en section 2.1). 44

descripteur une quantité mesurable ou calculable qui permet de décrire en partie une observation. Les MFCCs sont par exemple des descripteurs audio. Les descripteurs sont nommés *features* en anglais. 22–25, 32, 35, 36, 39, 47, 49, 52, 64, 70, 73, 85

grind inhale le *grind inhale* (GI) est une catégorie de distorsion vocale utilisée dans ce document pour désigner la technique vocale majoritairement utilisée dans le *grindcore*. Cri aspiré (voir description en section 2.1). 41, 43, 55, 74, 75

grindcore un des nombreux sous-genres extrêmes du *heavy metal*, ayant développé ses propres techniques vocales saturées (voir tableau 1.1). 11, 14, 16, 41–43, 82

hardcore scream le *hardcore scream* (HS) est une catégorie de distorsion vocale utilisée dans ce document pour désigner la technique vocale majoritairement utilisée dans le *metalcore*. Cri très saturé, particulièrement dans les fréquences aiguës (voir description en section 2.1). 41–43, 55, 57, 61, 64, 67, 73–76, 79–82

heavy metal également nommé plus simplement *metal*, le *heavy metal* est un genre musical inspiré du rock. Ce genre est devenu très populaire dans les années 1980, et a depuis développé de nombreux sous-genres (voir tableau 1.1). 11, 13, 14, 16, 41, 53, 80, 85

jitter irrégularité dans la période de la fréquence fondamentale. 21, 57

Machine Learning le *Machine Learning* est une branche de l'intelligence artificielle, qui explore les algorithmes pouvant améliorer leurs performances automatiquement à travers l'acquisition de données. 22–24, 29, 39, 42, 47, 49, 53, 55, 57, 85, 86

metalcore également nommé *hardcore*, il s'agit d'un des nombreux sous-genres extrêmes du *heavy metal*, ayant développé ses propres techniques vocales saturées (voir tableau 1.1). 11, 13, 14, 16, 81, 82, 85

observation donnée d'entrée de l'algorithme de *Machine Learning*. 22–25, 29, 31–36, 38, 39, 48, 52, 55, 64

pig squeal le *pig squeal* (PS) est un cri saturé imitant le son du cochon, utilisé comme effet (voir description en section 2.1). 43, 55, 74, 75

précision principal critère d'évaluation d'un algorithme de classification supervisée, il désigne le nombre d'observations correctement détectées dans la base de données de test, divisé par le nombre total d'observations de cette base de données. Lorsque ce terme est employé à propos d'un modèle de classification en classes multiples, il désigne la moyenne des précisions de chaque classe. La précision est nommée *accuracy* en anglais. 22, 23, 26, 29, 34, 38, 57, 64, 66, 67, 69, 70, 73, 75, 76, 79–82, 85

punk hardcore sous-genre du *rock*, inspiré du *punk* et étant considéré comme extrême dans ce style musical. Il a inspiré plusieurs sous-genres du *heavy metal*, tels que le *grindcore* et le *metalcore* (voir tableau 1.1). 13

shimmer variation d'amplitude dans la période de la fréquence fondamentale. 21, 57

thrash metal un des nombreux sous-genres du *heavy metal*, considéré comme extrême et qui a lui-même inspiré d'autres sous-genres extrêmes comme le *black metal* et le *death metal* (voir tableau 1.1). 13, 42

tunnel throat le *tunnel throat* (TT) est un *death growl* très grave, produit en recroquevillant la langue contre le palet, et utilisé comme effet (voir description en section 2.1). 44, 55, 74

undersampling méthode qui permet de rééquilibrer le nombre d'observations par classe, en supprimant un certain nombre d'observations appartenant à une ou plusieurs classes. 24, 55, 64, 73, 82

voix claire le terme voix claire, ou *clear voice* (CV) en anglais, désigne une voix non saturée. 18, 20, 31, 34, 44, 55, 57, 64, 67, 69, 70, 73, 74

Table des matières

Introduction **11**

1 État de l'art **13**

1.1 Histoire des techniques extrêmes de saturation vocale dans le *metal* 13

1.2 Taxonomies des différentes techniques de distorsion vocale 16

1.3 Production de la distorsion vocale 18

1.3.1 Production de la voix 18

1.3.2 Production des techniques de distorsion vocale 20

1.4 Études acoustiques des techniques extrêmes de distorsion vocale 21

1.5 Classification supervisée et Machine Learning 22

1.5.1 Undersampling et oversampling 24

1.5.2 Cross-validation 24

1.5.3 Descripteurs utilisés pour de la classification de signaux vocaux 25

Les Mel-Frequency Cepstal Coefficients (MFCC) 26

Le contraste spectral 28

1.5.4 Modèles de classification supervisée en classes multiples 29

Le perceptron multicouche 29

La forêt d'arbres décisionnels 34

La classification naïve bayésienne avec loi gaussienne 38

1.6 Conclusion pour l'état de l'art 39

2 Méthodes **41**

2.1 Taxonomie choisie pour la création de la base de données 41
2.1.1 Catégories de distorsion vocale 42
Le *black shriek* 42
Le *death growl* 42
Le *hardcore scream* 42
Le *grind inhale* 43
2.1.2 Effets de distorsion vocale 43
Le *pig squeal* 43
Le *deep gutturals* 44
Le *tunnel throat* 44
2.2 Méthodes d'enregistrement pour la création de la base de données 44
2.2.1 Matériel et lieu choisi pour les enregistrements 45
2.2.2 Préparation des enregistrements 45
2.2.3 Réalisation des enregistrements 46
2.2.4 Évaluation des enregistrements 46
2.3 Découpage des données, descripteurs utilisés et test de modèles de *Machine Learning* 47
2.3.1 Découpage des données 47
2.3.2 Extraction des descripteurs 47
Extraction des Mel Frequency Cepstral Coefficients (MFCC) 47
Extraction des Data Adjusted Frequencies Cepstral Coefficients (DAFCC) 47
Extraction du contraste spectral 52
2.3.3 Modèles de *Machine Learning* entrainés 52
Entraînement du perceptron multicouche 52
Entraînement de la forêt d'arbres décisionnels 52
Entraînement de la classification naïve bayésienne avec loi gaussienne . 53
2.4 Conclusion pour la partie Méthodes 53

3 Analyse des résultats **55**

3.1 Tri de la base de données 55

3.2 Analyse acoustique des différentes catégories 57

3.3 Analyse de plusieurs algorithmes : choix d'un modèle et de descripteurs 64

3.3.1 Influence d'un *undersampling* aléatoire sur la précision 64

3.3.2 Analyse comparative des modèles de *Machine Learning* 66

3.3.3 Suppression des silences 69

3.3.4 Analyse comparative des descripteurs 70

3.3.5 Modification de la taille des trames 72

3.4 Analyse du modèle final 73

3.4.1 Résultats globaux du modèle 73

3.4.2 Résultats par catégorie et registre 73

3.4.3 Résultats obtenus pour les techniques qui ne sont pas prises en compte par l'algorithme 74

3.4.4 Résultats par voyelles 75

3.4.5 Résultats par notes issues du tableau d'auto-évaluation 75

3.4.6 Résultats avec et sans tri 76

3.4.7 Résultats par sujets 76

3.4.8 Résultats par genre 79

3.5 Modèles proposés pour chaque sous-genre extrême du *heavy metal* 80

3.5.1 Modèle *death metal* 80

3.5.2 Modèle *black metal* 81

3.5.3 Modèle *metalcore* 81

3.5.4 Modèle *grindcore* (et *deathcore*) 82

3.6 Conclusion pour la partie analyse des résultats 82

Conclusion **85**

Bibliographie **87**

A Tableau d'auto-évalution et informations sur les chanteurs **95**

B Exemples proposés aux chanteurs lors des enregistrements **101**

Introduction

Les chanteurs de *heavy metal* sont très souvent amenés à utiliser des techniques de distorsion vocale pour changer la couleur de leur voix, ou pour en améliorer l'agressivité. Ces techniques sont produites en créant des comportements vibratoires inhabituels dans les cordes vocales et dans le conduit vocal, en faisant vibrer non seulement les cordes vocales, mais également d'autres parties du larynx comme les bandes ventriculaires. Les techniques employées sont diverses, et peuvent avoir un timbre et un mode de production très différents. Des techniques vocales assez distinctes sont notamment employées dans les sous-genres de *heavy metal* que sont le *death metal*, le *black metal*, le *metalcore*, et le *grindcore*.

Sur scène, un unique chanteur peut-être amené à alterner très souvent et très rapidement entre ces différentes techniques. Les traitements utilisés par les ingénieurs du son pour chacune d'elle peuvent être très variés, et sont parfois compliqués à mettre en place pour des performances live. Plusieurs dispositifs permettent aujourd'hui d'appliquer différents effets sur les différentes techniques vocales. Philippe Charny Dewandre, ancien chanteur du groupe *Kadinja*, utilise par exemple deux microphones différents sur scène en fonction des techniques utilisées. Certains chanteurs utilisent également des pédales d'effets. Ces dispositifs donnent cependant une charge supplémentaire au chanteur, ce qui peut nuire à sa performance. La charge pourrait revenir à l'ingénieur du son, qui pourrait envoyer le son dans des bus de traitements différents en connaissant la partition du chanteur, mais cela laisse alors peu de place à l'improvisation.

Une solution à ce problème consisterait à utiliser un plug-in capable de détecter automatiquement et en temps réel la technique vocale produite par le chanteur, avant de rediriger le signal dans des bus différents en fonction de la technique détectée. Les bus seraient donc munis d'un panel d'effets adaptés à la technique utilisée. Au cours de ce travail de mémoire, cette perspective sera abordée sous la forme d'une étude préliminaire, qui aura pour objectif de créer un algorithme de détection des différentes techniques vocales utilisées dans le *heavy metal*, algorithme susceptible d'être mis en œuvre en temps réel.

Dans le premier chapitre de ce document, un état de l'art sera abordé. Celui-ci présentera l'histoire de ces voix saturées dans le *metal* extrême, les différentes classifications existantes, ainsi que les algorithmes de classification ayant été développés dans des projets similaires. Dans le second chapitre, une nouvelle classification de ces techniques sera proposée, et servira de repère pour la création d'une base de données alimentant un algorithme de classification supervisée. Les performances de cet algorithme seront enfin analysées dans le troisième chapitre de ce document.

Chapitre 1

État de l'art

1.1 Histoire des techniques extrêmes de saturation vocale dans le *metal*

L'histoire des techniques vocales extrêmes dans le *metal* est intrinsèquement liée à l'émergence de différents sous-genres extrêmes du *heavy metal* (voir figure 1.1). Le *black metal*, influencé par le *thrash metal*, naît avec le groupe *Venom* et leur album *Black Metal (1982)*. Les techniques vocales du *black metal* se caractérisent assez rapidement par des couleurs vocales très aiguës et saturées, des sortes de cris perçants, que l'on peut trouver chez certains chanteurs précurseurs comme Quorthon du groupe *Bathory*. Le style vocal se perfectionne, avec des groupes comme *Sarcofago* et leur opus *INRI (1987)*, et sera devenu dès lors la technique vocale la plus répandue pour ce genre musical. Le *black metal* gagnera énormément en popularité au début des années 1990 grâce à des groupes norvégiens tels que *Darkthrone*, *Dimmu Borgir* ou *Immortal*.

Le *death metal*, également inspiré du *thrash metal* et étroitement lié au *black metal*, est un sous-genre qui a développé ses propres techniques vocales gutturales. Les premiers groupes de *metal* à assumer dans leur esthétique des chants gutturaux, seraient *Mantas* dans leur album démo *Death By Metal (1984)* ou encore *Possessed* dans leur album démo nommé *Death Metal (1984)*. Purcell (2003) décrira les techniques employées par Jeff Beccera, le chanteur de *Possessed*, comme étant des « grognements bien plus brutaux que les voix les plus extrêmes du *trash metal* ». Ces techniques seront alors popularisées par des groupes de *death metal* tels que *Morbid Angel* et *Sepultura*, vers la fin des années 1980. Des groupes tels que *Cannibal Corpse* viendront perfectionner cette technique vocale dans les années 1990.

Vers la fin des années 1980 et dans les années 1990, le *heavy metal* se diversifie, et de nombreux nouveaux sous-genres émergent, comme le *metalcore*. Le *metalcore*, également nommé *hardcore*, est un genre inspiré du *punk hardcore*, au sein duquel les chanteurs ont commencé à utiliser des techniques vocales de plus en plus extrêmes. Une technique de voix saturée, communément appelée *screaming*, devient une norme de ce nouveau genre

musical, avec des groupes comme *Converge* ou *Earth Crisis*. Cette technique est particulièrement perfectionnée par Greg Puciato de *The Dillinger Escape Plan* à la fin des années 1990. Ce genre reste aujourd'hui un des genres les plus populaires du *metal*, avec notamment le groupe *Bring Me The Horizon* et son chanteur Oliver Sykes qui utilise principalement des techniques de chant clair et de *screaming*. La technique du *screaming* émerge aussi dans d'autres sous-genres du *heavy metal* dans les années 1990, comme dans le *deathcore*, le *nu metal*, le *metal alternatif*, le *metal industriel* ou dans le *grindcore*.

D'autres techniques vocales, notamment des cris produits en inhalant l'air plutôt qu'en l'expirant, émergent dans le *grindcore* et dans le *death metal* vers la fin des années 1990. S'il est difficile de retracer les premières utilisations des cris inhalés, il est possible cependant de remarquer leur utilisation particulièrement proéminente par plusieurs chanteurs de *grindcore*, comme le chanteur de *Sublime Cadaveric Decomposition*.

Les techniques vocales citées précédemment, bien que presque toutes rattachées à un genre musical unique, n'y sont pas forcément cantonnées. Hainaut (2020) dira par exemple que le *death metal* étant un style plus technique que le *black metal*, les chanteurs y utiliseront plus souvent des techniques vocales plus diversifiées, et pourront même y utiliser des techniques propres au *black metal*. C'est le cas par exemple dans la chanson *Blunt Force Trauma* [1] (issu de l'album Purification Through Violence (1996)) de *Dying Fetus* où John Gallagher alterne entre des voix propres au style *death metal* et des voix plutôt typées *black metal*. Dans le *black metal*, certains chanteurs, comme Darren White du groupe *Cradle Of Filth*, utilisent parfois en plus des techniques vocales de *black metal*, des cris identiques aux techniques utilisées dans le *metalcore*. De nos jours, de nombreux chanteurs exploitent sur scène ces différentes techniques vocales, qu'elles proviennent du *hardcore*, du *death metal*, du *black metal* ou du *grindcore*. Julien Truchan du groupe *Benighted*, mêle par exemple l'utilisation de toutes ces techniques dans *Reptilian* (issu de l'album Necrobreed (2017)).

1. cliquer sur le nom de la chanson pour accéder à un extrait musical. D'autres chansons citées dans ce document seront accompagnées d'extraits musicaux

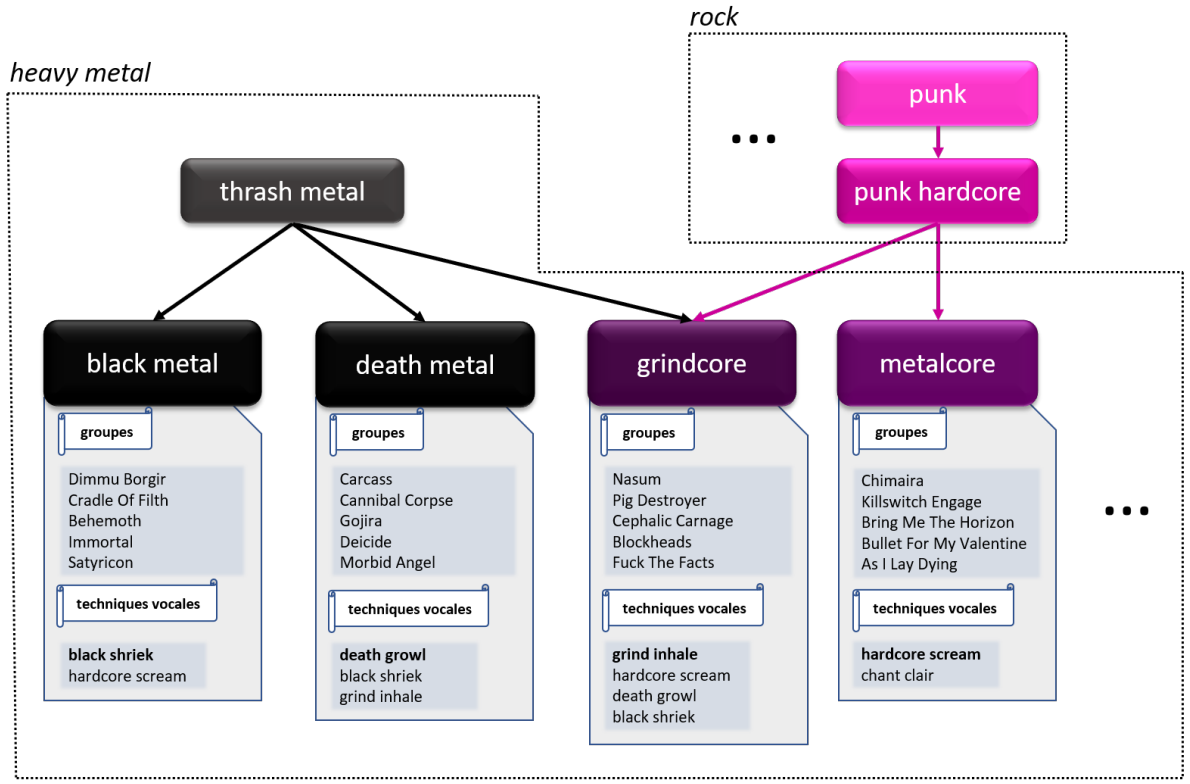

Figure 1.1 – Schéma représentant les influences des sous-genres les plus extrêmes du *heavy metal*. Les techniques vocales citées dans ce schéma proviennent de la classification établie en section 2.1. Les flèches représentent les influences d'un genre sur un autre. Ici, seuls les sous-genres extrêmes du *heavy metal* ayant été particulièrement innovants dans l'utilisation de techniques vocales saturées ont été représentés. Cependant, bien d'autres sous-genres extrêmes existent (comme le *deathcore*, le *nu metal*, le *metal industriel*, le *metal alternatif*, etc.) et beaucoup de groupes de musique appartenant à ces sous-genres pratiquent ces mêmes techniques .

1.2 Taxonomies des différentes techniques de distorsion vocale

Ce travail se focalise essentiellement sur les techniques de distorsion vocale les plus extrêmes ayant été étudiées, en excluant les techniques de distorsion plus légères comme le *rattle* (employé par des chanteurs comme Joe Cocker), le *growl* (employé par Louis Armstrong), ou la *légère distorsion* (employée par Kurt Cobain) (voir taxonomie de Sadolin (2000)). Pour simplifier cette approche, les techniques de saturation plus légères sont considérées comme une simple variation dans la prosodie du chant clair, du fait que leur utilisation ne requiert pas un mixage particulier de la part de l'ingénieur du son lors de performances live. Il est à noter ici que les techniques vocales extrêmes employées dans le *heavy metal* ont été jusqu'alors classifiées de façon très empiriques, et que les différents chercheurs ayant travaillé sur le sujet peuvent de ce fait être en désaccord sur la nomenclature à utiliser.

Nieto (2013, 2008) a établi une taxonomie des techniques vocales les plus extrêmes en trois catégories :

— Le *death growl* : cette technique est très utilisée dans le *death metal*. La voix résultante est très bruitée, avec un ajout particulièrement prononcé de fréquences inférieures à la fréquence fondamentale.

— Le *fry scream exhale* : particulièrement utilisée dans le genre *metalcore*, cette technique possède un timbre plus clair que le *death growl*.

— Le *fry scream inhale* : cette technique trouve particulièrement sa place dans le *grindcore* et le *deathcore*. Elle peut également être nommée *pig squeal* du fait de sa ressemblance avec le cri du cochon.

Sadolin (2000) ne donne des exemples issus du genre métal que pour la technique *grunt* (qui est un synonyme du *death growl*) dans son ouvrage. D'après elle, cette technique est particulièrement utilisée dans le *death metal* et le *black metal*, et elle la définit comme étant un mélange de *growl* et de légère distorsion. Elle décrit le son résultant comme étant grave, sombre et caverneux. Sans se référer explicitement au *metal*, mais en citant par exemple le *hard rock* ou le *punk*, elle mentionne également une technique nommée *distorted scream*, et qui se rapproche de la description du *fry scream* de la taxonomie de Nieto (2008, 2013).

Lors d'une étude récente, Hainaut (2020) constate des différences significatives entre les voix utilisées dans le *death metal*, et celles utilisées dans le *black metal*. Si dans les deux genres, les chanteurs utilisent la même technique vocale (qui serait un *death growl* d'après les taxonomies précédentes), ils ne l'utilisent pourtant pas exactement de la même façon. Hainaut a pu estimer que les hauteurs employées pour cette technique dans le *death metal* et le *black metal* sont significativement différentes : la voix *black* est employée dans un registre plus aigu que la voix *death* et elle possède un timbre plus clair et une utilisation différente du voisement.

Purcell (2003) définit un autre type de voix métal, issue du *death metal* new-yorkais, qu'elle qualifie de « sépulcrales ». Ces voix sont des voix de *death metal* encore plus graves

que ce qui est habituellement produit, les paroles des chansons devenant quasiment inintelligibles. Elle cite par exemple le groupe *Dying Fœtus*, avec le chanteur John Gallagher qui produit régulièrement cette technique comme il peut le faire dans le morceau *Absolute Defiance*.

Chevaillier *et al.* (2011) ont établi une taxonomie des voix utilisées dans le métal en quatre catégories, en suivant les dénominations employées par un professeur de chant : la saturation nasale, la saturation vélaire, la saturation supra-glottique et la saturation glotto-vélaire. Leur panel d'auditeurs experts, composé d'orthophonistes, de musiciens et d'acousticiens, n'a réussi à déceler que deux catégories distinctes dans ce corpus, séparant assez clairement les voix issues de la saturation nasale et vélaire, et celles issues de la saturation supra-glottique et glotto-vélaire.

Les différentes dénominations ont été mises en correspondance dans le tableau 1.2, en comparant les exemples et les genres musicaux donnés par chacune des recherches. L'absence d'exemples musicaux dans l'étude de Chevaillier *et al.* (2011) ne permet pas de mettre en correspondance les techniques catégorisées avec celles étudiées ou définies dans les autres recherches.

Catégories	Nieto (2008, 2013)	Sadolin (2010)	Hainaut (2020)	Thuesen (2017)	Purcell (2003)	Chevaillier (2008)
death growl	death growl	grunt	voix death metal	grunt	voix death metal	
black shriek			voix black metal			
hardcore scream	fry scream exhale	distorted scream				
grind inhale	fry scream inhale					
deep gutturals					voix sépulcrale	
						saturation nasale et vélaire
						saturation suppraglotique et glottovélaire

Figure 1.2 – Comparaison des différentes taxonomies de voix extrêmes saturées. Les catégories présentées dans la première colonne sont celles qui ont été choisies dans cette étude (voir partie 2.1).

1.3 Production de la distorsion vocale

1.3.1 Production de la voix

Dans la production d'une voix claire, c'est-à-dire une voix non saturée, un flux d'air provenant des poumons va traverser les cordes vocales (situées dans le larynx) afin de générer un son. Ce son est appelé source vocale. La source vocale passe ensuite dans le conduit vocal, dans lequel elle est modifiée acoustiquement. Sundberg et Rossing (1987) comparent les poumons à un compresseur, les cordes vocales à un oscillateur, et le conduit vocal à un résonateur (voir figure 1.3).

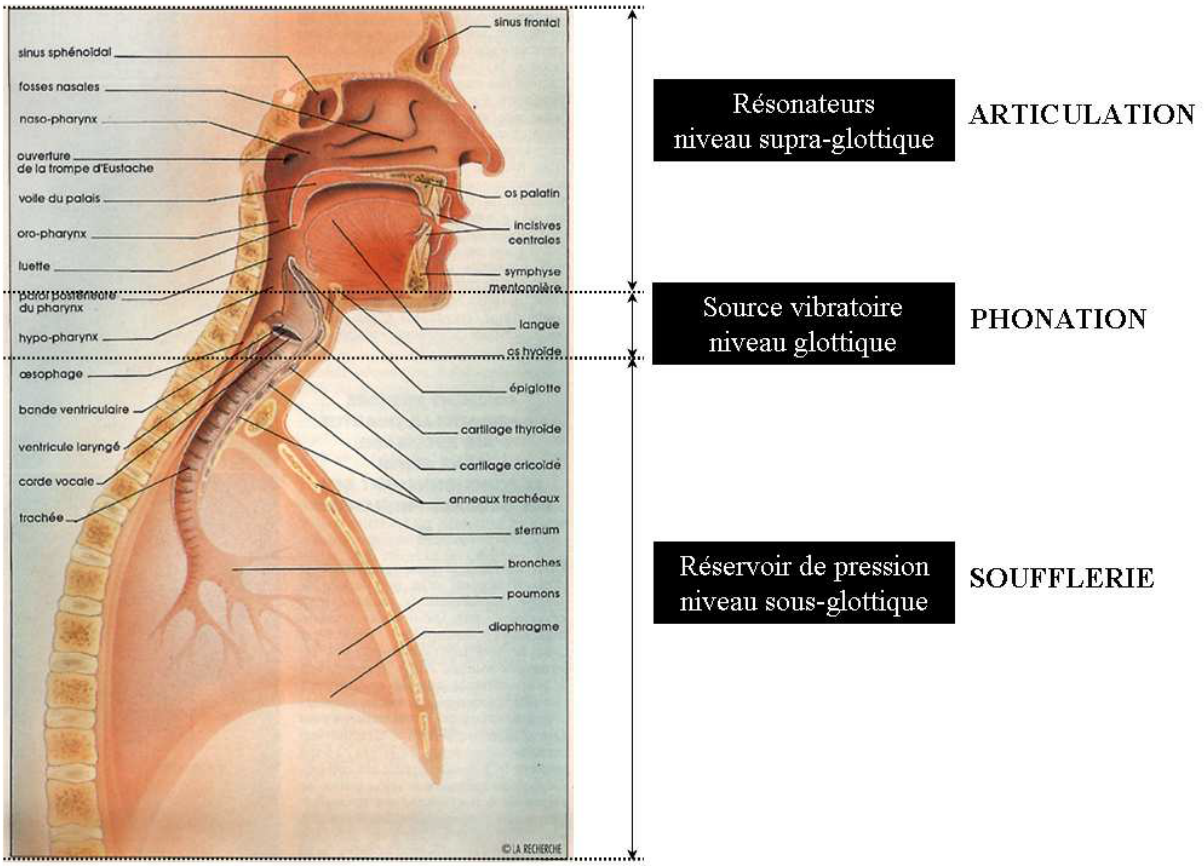

Figure 1.3 – Appareil vocal humain, d'après Scotto Di Carlo (1991)

.

Afin de comprendre comment est produite la distorsion vocale, il est important d'analyser les différents cartilages du larynx. Le larynx est constitué d'un ensemble de pièces cartilagineuses, reliées entre elles par des ligaments, et le tout mis en mouvement par des muscles, des muqueuses, des cavités, des membranes, des vaisseaux et des nerfs (Bailly, 2009) (voir figure 1.4).

Le larynx est notamment composé des cartilages suivants :

— Le cricoïde : dernier anneau de la trachée,

— Le thyroïde : lame en angle dièdre, sur laquelle les cordes vocales viennent prendre appui,

— L'épiglotte : lame oblique qui protège partiellement la cavité laryngée,

— Les aryténoïdes : cartilages sur lesquelles s'insèrent chacune des deux cordes vocales. Ils peuvent se mouvoir très rapidement, ouvrant ou fermant la glotte, ce qui permet de séparer ou de rejoindre les parties supérieures des cordes vocales,

— Les cartilages corniculés : cartilages situés au-dessus des aryténoïdes,

— Les cartilages cunéiformes : cartilages situés dans les replis aryténo-épiglottiques.

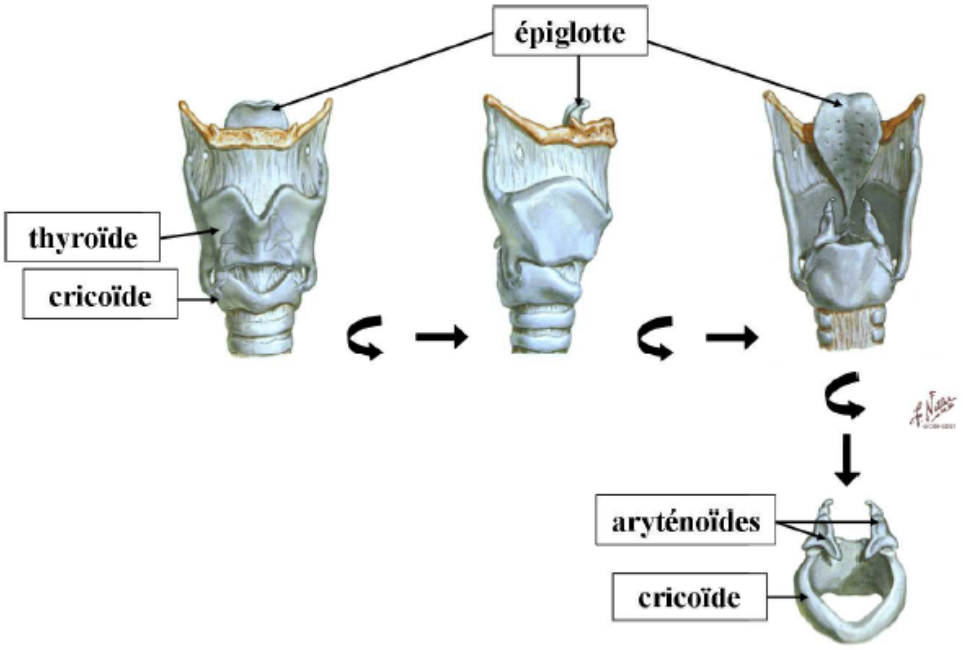

Figure 1.4 – Cartilages du larynx, d'après Netter et Scott (2007)

Il existe également deux replis muqueux au sein du larynx :

— Les cordes vocales : le locuteur peut en contrôler plusieurs paramètres, comme l'écartement, la longueur, la forme et la tension, en contractant les muscles du larynx, ou en modifiant la position des différents cartilages du larynx. La fente entre les cordes vocales est nommée la glotte.

— Les bandes ventriculaires : elles se situent juste au-dessus des cordes vocales, possèdent des propriétés similaires, et sont donc souvent également dénommées « fausses cordes ».

La glotte se situe à l'extrémité inférieure du tube laryngé, qui s'insère dans un tube plus large et plus grand nommé le pharynx. À droite et à gauche du tube laryngé, la partie inférieure du pharynx forme des orifices appelés les orifices piriformes. L'hypopharynx, qui se trouve derrière l'orifice piriforme, referme le larynx lors de la déglutition. La combinaison du pharynx et de la cavité buccale est nommée le conduit vocal.

Dans le cas de la production d'une voix saturée, les différents cartilages du larynx, qui servent à soutenir les cordes vocales pour produire une voix claire, servent eux-mêmes ici d'oscillateur au même titre que les cordes vocales, et vont donc entrer en vibration lors du passage du flux d'air. Les bandes ventriculaires sont également mises en jeu. Une partie du pharynx, et donc une partie du résonateur de la voix claire, pourront aussi servir d'oscillateur.

1.3.2 Production des techniques de distorsion vocale

Différentes études ont été menées afin d'étudier plus en détail les différentes parties du larynx pouvant entrer en vibration lors de l'utilisation de diverses techniques de distorsion vocale. Lors de leurs différentes études utilisant une laryngo-stroboscopie, McGlashan *et al.* (2013, 2017) et Thuesen *et al.* (2017), ont établi un découpage du conduit vocal en 6 niveaux (voir figure 1.5) :

— Niveau 1 : cordes vocales,

— Niveau 2 : bandes ventriculaires,

— Niveau 3 : cartilages aryténoïdes et cunéiformes, épiglotte, et plis ary-épiglottiques,

— Niveau 4 : orifice piriforme et paroi pharyngée postérieure de l'hypopharynx,

— Niveau 5 : palais mou, luette, paroi arrière de la gorge (oropharynx) et dos de la langue,

— Niveau 6 : reste du conduit vocal (cavités buccales et nasales).

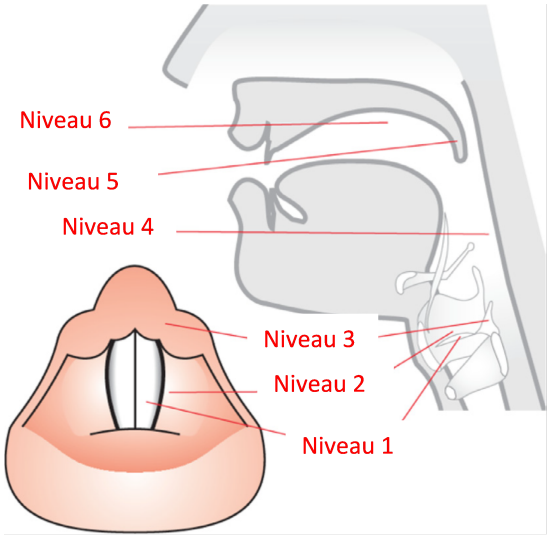

Figure 1.5 – Les 6 niveaux du conduit vocal, traduit de McGlashan *et al.* (2017).

Le *grunt* (ou *death growl*), d'après Thuesen *et al.* (2017), est produit en faisant vibrer toute la structure supra-glottique (c'est-à-dire au-dessus de la glotte) des niveaux 1 à 4. De grandes amplitudes de vibration des cordes vocales ont été observées, ainsi que des mouvements particuliers des plis ary-épiglottiques. La technique de *rattle* est produite en faisant

vibrer les cartilages arytenoïdiens l'un contre l'autre. Le *growl*, quant à lui, serait produit en faisant vibrer les cartilages arytenoïdiens contre l'épiglotte, ou proche de l'épiglotte. Enfin, la technique nommée distorsion est produite en faisant vibrer les plis ventriculaires l'un contre l'autre. Le caractère extrême de la technique de *grunt* (*death growl*) est ici mis en évidence, car contrairement aux trois autres techniques plus légères, ce n'est pas seulement une partie de la structure supra-glottique qui se met à vibrer lors de sa production, mais son intégralité. Cependant, le niveau de vibration des différentes structures lors de l'utilisation de cette technique varie beaucoup en fonction du chanteur.

Chevaillier *et al.* (2011), constatent une vibration des bandes ventriculaires pour chacune des techniques qu'ils ont définies, c'est-à-dire pour la saturation nasale, vélaire, supra-glottique et glotto-vélaire.

D'après Nieto (2013), le *fry scream* serait produit par une série de pulsations glottales espacées dans le temps (fonctionnant de la même manière pour le *fry scream inhale* et *exhale*). Cependant, contrairement aux études effectuées sur les techniques citées précédemment (Chevaillier *et al.*, 2011; Thuesen *et al.*, 2017), ces constatations sur la vibration des différentes parties du larynx lors de la production du *fry scream* ne sont pas appuyées par une étude laryngo-stroboscopique, ce qui limite la pertinence de cette remarque.

1.4 Études acoustiques des techniques extrêmes de distorsion vocale

Dans le domaine spectral, la distorsion vocale se caractérise par un rapport harmoniques-bruit (HNR) très faible (Yumoto *et al.*, 1984; Tsai *et al.*, 2010), et par la présence de bruit et de sous-harmoniques (Fitch *et al.*, 2002; Sakakibara *et al.*, 2004; Nieto, 2008; Bonada et Blaauw, 2013). Le terme sous-harmonique décrit ici des signaux dont la fréquence est une division de la fréquence fondamentale f0 par un nombre entier (f0/n avec n=1,2,3,4...), mais également les composantes harmoniques des nouvelles sous-harmoniques créées (mfo/n avec n=1,2,3,4... et m=1,2,3,4...) (Loscos et Bonada, 2004; Gentilucci *et al.*, 2019). Nieto (2008) constate également l'émergence de sous-harmoniques instables et une certaine quantité de bruit dans le son pour les techniques de distorsion vocale les plus légères. Dans les distorsions vocales les plus extrêmes, le son devient totalement bruité et la fréquence fondamentale du signal devient inidentifiable (Nieto, 2008; Smialek *et al.*, 2012).

Dans le domaine temporel, elle est caractérisée par la présence significative de *jitter* et de *shimmer* dans la voix (Jones *et al.*, 2001; Verma et Kumar, 2005; Bonada et Blaauw, 2013). Le *jitter* peut être défini comme une irrégularité dans la période de la fréquence fondamentale, c'est-à-dire que chaque cycle de cette période peut avoir une durée légèrement différente. Le *shimmer* décrit une variation d'amplitude des différentes périodes, c'est-à-dire que chaque cycle de cette période aura une amplitude différente (Gentilucci *et al.*, 2019). Pour certains types de saturation vocale, des macro impulsions sont observables. Les macro impulsions sont des groupes d'impulsions glottales à forme et amplitude variables, possédant une fréquence inférieure à la fréquence fondamentale (Gentilucci *et al.*,

2019) (voir figure 1.6). L'impulsion glottale correspond à une impulsion parmi le train d'impulsions lié au jet d'air produit lors de l'ouverture des cordes vocales. Les distorsions vocales sont plus ou moins stables, et plusieurs d'entre elles, comme le *screaming* (*hardcore scream* selon la taxonomie établie en section 2.1), peuvent présenter des bifurcations. Les bifurcations sont des transitions soudaines et incontrôlées entre différents comportements vibratoires, c'est-à-dire la variation du nombre de sous-harmoniques présentes dans le signal.

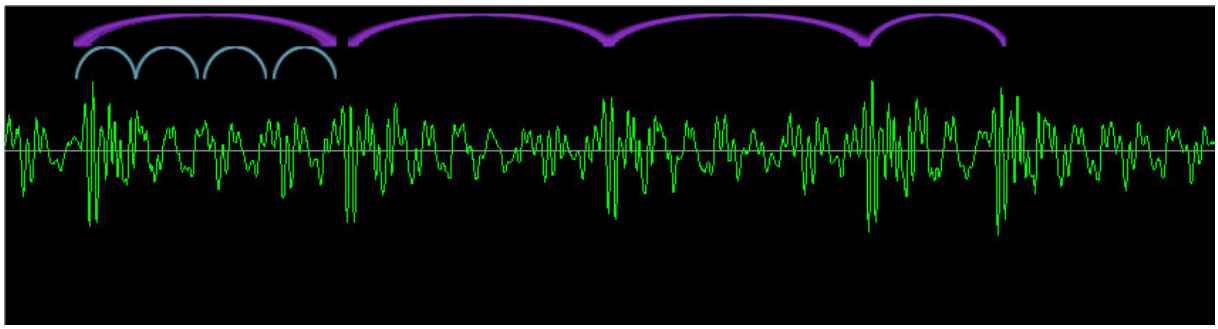

Figure 1.6 – Quatre macro-impulsions (en violet) qui regroupent plusieurs impulsions glottales (en bleu) pour un signal de distorsion vocale (illustration issue de Nieto (2008)).

1.5 Classification supervisée et Machine Learning

Le *Machine Learning* (également nommé apprentissage automatique) est une branche de l'intelligence artificielle, qui explore les algorithmes pouvant améliorer leurs performances automatiquement à travers l'acquisition de données. Le développement d'un algorithme de *Machine Learning* s'effectue en deux phases : une phase d'apprentissage (ou d'entraînement), lors de laquelle l'algorithme va extraire plusieurs descripteurs des données et entraîner un modèle, et une phase de test, lors de laquelle les performances de l'algorithme seront évaluées. Chaque nouvelle donnée apportée au modèle, que ce soit pour l'entraînement ou pour le test, sera dénommée observation.

La qualité des observations, leur nombre, et les métadonnées qui les accompagnent vont orienter la méthode employée pour créer un algorithme de *Machine Learning*. Les algorithmes en classification supervisée sont utilisés lorsque les données sont déjà étiquetées, c'est-à-dire que la réponse attendue de l'algorithme est déjà connue. Lorsque les données sont étiquetées au sein d'un ensemble de plus de deux classes, ce qui sera le cas dans ce travail, la classification sera nommée classification en classes multiples.

La phase de test de la classification supervisée s'effectue en plusieurs étapes. La première étape est une étape de paramétrisation, qui consiste à extraire du signal un ou plusieurs descripteurs acoustiques (également appelés paramètres acoustiques). Un modèle de *Machine Learning* entraîné sert à émettre des prédictions sur la classe de nouvelles observations (voir figure 1.7). La précision d'un algorithme désigne le nombre de prédictions correctement identifiées dans la base de données de test, divisé par le nombre total

de prédictions de cette base de données. Dans cette étude, le terme précision, quand il ne sera pas associé à une catégorie en particulier, désignera la précision moyenne entre les catégories.

Le modèle de *Machine Learning* est entraîné à partir d'une base de données d'entraînement, qui peut être différente de la base de données de test. L'extraction des descripteurs audio permet ici d'entraîner le modèle avec chaque observation en lui permettant de s'adapter en fonction de la classe attendue afin de générer la précision la plus élevée possible (voir figure 1.7).

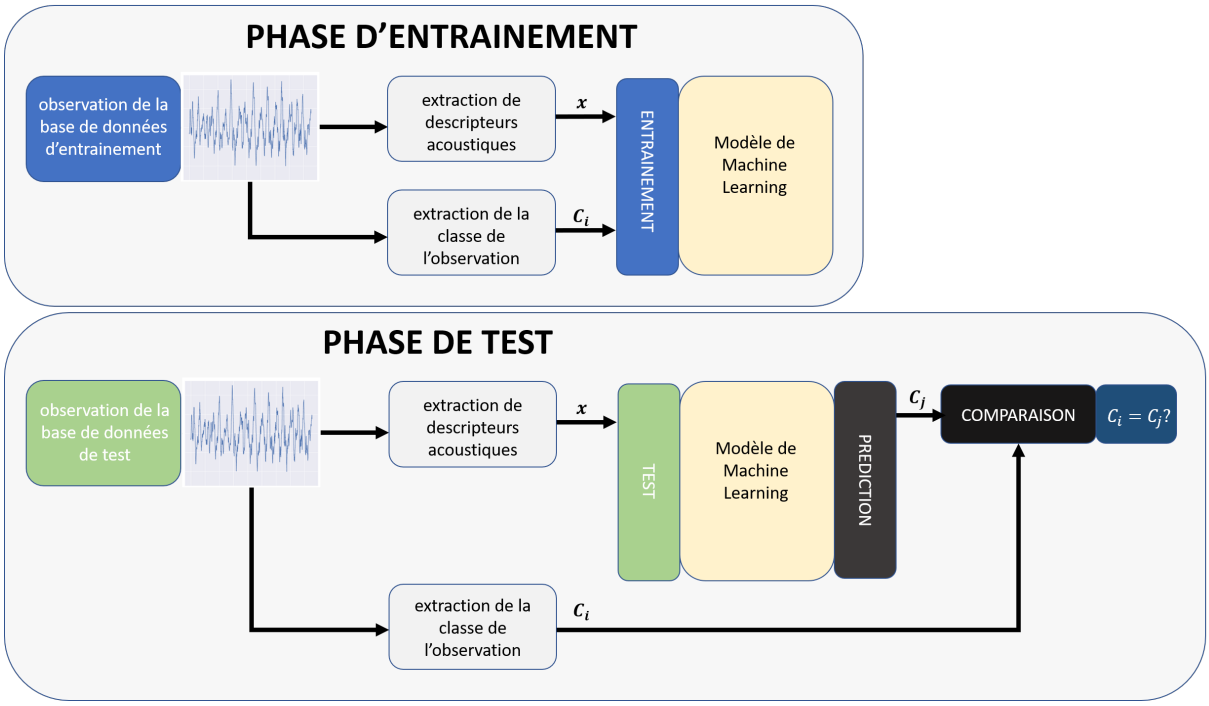

Figure 1.7 – Principe général pour l'entraînement et le test d'un modèle de *Machine Learning* en classification supervisée. C_i désigne ici la classe i de l'observation. C_j désigne la classe prédite j par l'algorithme.

Nieto (2013) a travaillé sur un modèle de *Machine Learning* permettant d'attribuer une classe à différentes voix saturées. Bien que très proche de l'étude de Nieto, la présente étude s'en éloigne pour plusieurs raisons :

— La base de données consolidée dans ce projet contient uniquement des voix non traitées et beaucoup moins bruitées (la base de données de Nieto était composée d'extraits musicaux dans lesquels les voix étaient traitées et mixées avec les autres instruments).

— un tel algorithme doit ici pouvoir fonctionner en temps réel, et la fenêtre d'étude doit donc être très petite.

— les données de l'étude sont déjà étiquetées, autrement dit la classification est supervisée.

Les parties 1.5.3 et 1.5.4 présentent les descripteurs et les modèles de *Machine Learning* les plus utilisés dans la classification de signaux sonores, et plus particulièrement de signaux vocaux.

1.5.1 Undersampling et oversampling

Lorsque le nombre d'observations est déséquilibré entre les différentes classes (c'est-à-dire que le ratio des différentes classes est significativement différent), il peut être difficile d'entraîner un modèle de *Machine Learning*. En effet, l'abondance d'observations pour une classe peut amener le modèle à être sur-entraîné pour cette classe, et il pourrait ainsi négliger son entraînement pour les autres classes. Une solution à ce problème est d'avoir recours à des techniques d'*undersampling* ou d'*oversampling*. L'objectif de chacune des deux techniques est de rééquilibrer le ratio des différentes classes.

Une première méthode très simple d'*undersampling* consiste à sélectionner aléatoirement un nombre d'observations à supprimer parmi la ou les classes dominantes. Cette technique, bien que communément utilisée, peut avoir un effet sur la variance de la base de données et des informations importantes peuvent être perdues. Bien d'autres méthodes existent, comme celle qui consiste à remplacer des groupes de données d'une même classe par les moyennes des K clusters calculés via un algorithme de type *K-moyennes*[2], où le nombre de clusters représente le nombre final d'observations (se référer aux travaux de Lin *et al.* (2017) pour obtenir plus d'informations sur cette technique). Cependant, Dal Pozzolo *et al.* (2015) mettent en garde sur l'utilisation naïve et systématique de l'*undersampling*, il conviendra donc de comparer les résultats avec et sans l'utilisation de cette méthode.

Comme pour l'*undersampling*, une utilisation simple de l'*oversampling* peut être la copie aléatoire d'une partie ou de la totalité des observations parmi la ou les classes dominées. Bien d'autres méthodes d'*oversampling* ont pu prouver leur efficacité, comme l'algorithme SMOTE (Fernández *et al.*, 2018) ou ADASYN (He *et al.*, 2008).

1.5.2 Cross-validation

La *cross-validation* désigne l'ensemble des techniques permettant de séparer les données d'entraînement des données de test, afin de s'assurer de tester un modèle sur des données n'ayant jamais servi pour son entraînement.

En particulier, dans le cas de l'enregistrement de chanteurs, la *cross-validation* en *N-folds* consiste à garder un unique chanteur pour la phase de test et entraîner le modèle sur les données de tous les autres chanteurs. Les prédictions obtenues pour les observations de ce chanteur seront conservées, et l'opération sera répétée pour chacun des chanteurs. L'algorithme créera ainsi autant de modèles qu'il y a de chanteurs. Les prédictions seront

2. Les algorithmes de K-moyennes permettent de répartir les données dans les K meilleurs groupes (aussi appelés *clusters*) sur la base de certains critères, et ne gardent que la moyenne des données associées à chaque groupe

ensuite regroupées pour former le vecteur de prédiction total. Cette méthode permet de bien séparer les données d'entraînement et de test, tout en testant l'algorithme sur l'ensemble des observations (voir figure 1.8).

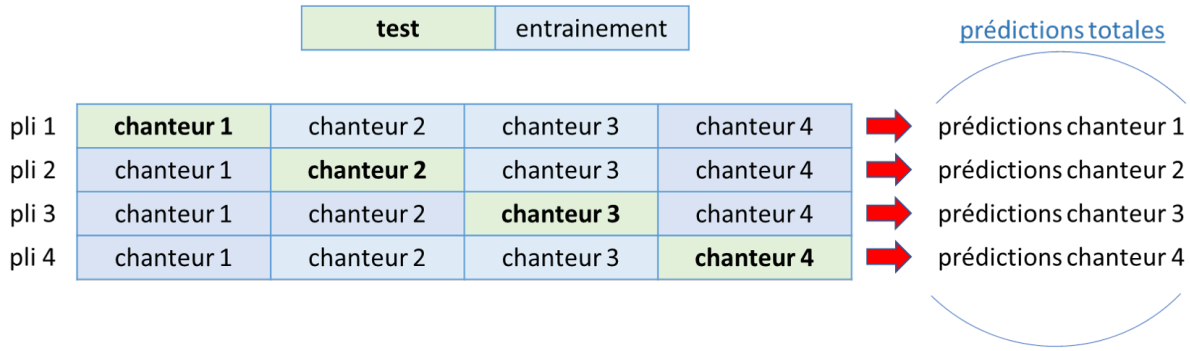

Figure 1.8 – Cross-validation en *N-folds* pour une base de données contenant les enregistrements de quatre chanteurs.

1.5.3 Descripteurs utilisés pour de la classification de signaux vocaux

L'objectif de ce travail est de répartir des données de voix dans différentes classes en fonction de leur timbre (et non en fonction de leur hauteur, de leur intensité ou de leur spatialisation). Par conséquent, ne seront présentés ici que les descripteurs ayant été particulièrement performants dans l'analyse du timbre de la voix.

D'après Rocamora et Herrera (2007) les Mel Frequency Cepstral Coefficients (MFCC), les Linear Prediction Coefficients (LPC)[3], les Log Frequency Power Coefficients (LFPC)[4], le flux spectral[5], les centroïdes spectraux[6] ainsi que les roll-off spectraux[7] ont montré de très bons résultats pour la détection de signaux de voix basée sur le timbre. Mesaros *et al.* (2017), décrivent les MFCC comme étant les descripteurs les plus utilisés pour la détection d'évènements sonores. Rocamora et Herrera (2007), après avoir comparé plusieurs descripteurs pour de la classification de voix, concluent qu'ils sont les paramètres les plus performants pour leur étude. Nieto (2013), lors de son étude sur la détection de voix saturées, a par exemple extrait plusieurs descripteurs, dont les 13 premiers MFCC.

3. les LPC permettent de récupérer des informations sur l'enveloppe spectrale du signal sous une forme compressée, en utilisant les informations d'un modèle prédictif linéaire
4. Les LFPC sont obtenus en récupérant l'énergie de chaque filtre issus d'une banque de filtres logarithmique entre 200 Hz et 4 kHz (Nwe *et al.*, 2004)
5. Le flux spectral est une mesure de la variation du spectre d'un signal, en comparant le spectre en puissance d'une trame et celui de la trame précédente
6. Le centroïde spectral d'une trame correspond à la fréquence pour laquelle l'énergie du spectre est équitablement répartie de part et d'autre
7. Le roll-off spectral d'une trame correspond à la fréquence pour laquelle x% de l'énergie du spectre est inférieure. x est souvent fixé à 85

De nombreux travaux reposent ainsi uniquement sur l'extraction des MFCC pour classifier leurs données, comme l'étude de Iheme et Ozan (2019). Alsouda *et al.* (2018) ont obtenu une précision variant entre 85% et 100% pour de la classification de bruits. Le contraste spectral est également un descripteur assez courament utilisé, et est exploité par exemple dans les travaux de Nieto (2013).

Les Mel-Frequency Cepstal Coefficients (MFCC)

Les MFCC sont les résultats d'une transformation en cosinus du logarithme du spectre de puissance du signal, exprimé sur une échelle de fréquences en mels. La méthode employée pour extraire les MFCC est notamment détaillée dans les travaux de Oo (2018). Ces coefficients sont calculés à l'aide de transformées de Fourier à court-terme, calculées pour des trames temporelles assez petites (par exemple de 1024 échantillons).

Tout d'abord, le signal est divisé en trames de petite taille. Puis un fenêtrage est appliqué, par exemple avec une fenêtre de Hann. Ensuite, une transformée de Fourier rapide est appliquée (voir équation 1.1).

$$S_i(k) = \sum_{n=1}^{N} s_i(n) w(n) e^{-j2\pi kn/N} \tag{1.1}$$

$s_i(n)$ désigne le signal pour la $i^{\text{ième}}$ trame, et $S_i(k)$ le résultat de la transformée de Fourier rapide pour la $k^{\text{ième}}$ fréquence de la $i^{\text{ième}}$ trame après fenêtrage. N représente ici le nombre d'échantillons par trame, et w la fenêtre d'analyse choisie. k, entier compris entre 1 et N, correspond à l'indice de la fréquence étudiée.

À l'issue de cette transformée de Fourier, ne seront conservées que les K premières fréquences car les informations associées aux fréquences suivantes correspondent aux fréquences négatives, qui n'apportent pas plus d'informations du fait de la symétrie hermitienne des transformées de Fourier. K prend la valeur définie en équation 1.2.

$$K = \frac{N}{2} + 1 \tag{1.2}$$

Le périodogramme est ensuite obtenu en effectuant le calcul présenté en équation 1.3, où $P_i(k)$ désigne la puissance spectrale de la $k^{\text{ième}}$ fréquence de la $i^{\text{ième}}$ trame. La puissance totale d'une trame i désignera la somme de toutes les puissances de cette trame (voir équation 1.4)

$$P_i(k) = \frac{1}{K} |S_i(k)|^2 \tag{1.3}$$

$$P_i = \sum_{k=1}^{K} P_i(k) \tag{1.4}$$

Une banque de filtres de Mel est alors générée. Cette banque de filtres est construite à partir de plusieurs filtres passe-bande, sous la forme de fenêtres triangulaires (Davis et

Mermelstein, 1980). Les largeurs de bande sont fixées de manière à obtenir un chevauchement de 50% entre les fenêtres. Les fréquences centrales en Hz sont déterminées à partir de l'échelle de Mel, et sont toutes réparties uniformément sur cette échelle. L'échelle de Mel est une échelle perceptuelle établie à partir d'études sur la largeur des bandes critiques de l'oreille (Picone, 1993).

Il existe plusieurs formules alternatives permettant de mettre en correspondance les fréquences en hertz et les fréquences en mels. Slaney (1998) propose d'utiliser une échelle de conversion linéaire jusqu'à 1 kHz (voir équation 1.5) , puis logarithmique à partir de 1 kHz (voir équation 1.6). Dans ces deux équations $f_{sp} = 200/3$ et $logstep = ln(6,4)/27$ sont des facteurs d'échelle, et f_{min} désigne la fréquence minimale étudiée. f représente alors la fréquence en hertz, et f_{mel} la fréquence correspondante en mels. Cette formule sera nommée formule de Slaney.

$$\underset{f \geq 1000Hz}{f_{mel}} = \frac{(1000 - f_{min})}{f_{sp}} + ln(\frac{f}{1000})\frac{1}{logstep} \tag{1.5}$$

$$\underset{f < 1000Hz}{f_{mel}} = \frac{(f - f_{min})}{f_{sp}} \tag{1.6}$$

Young *et al.* (2002), quant à eux, proposent une autre formule pour convertir les fréquences en mels (voir équation 1.7). D'après Ganchev *et al.* (2005), cette formule permet d'avoir une meilleure approximation de l'échelle de Mel pour les fréquences en dessous de 1 kHz, mais est moins précise pour les fréquences supérieures à 1 kHz. Cette formule sera nommée formule de Young.

$$f_{mel} = 2595 \cdot \log_{10}(1 + \frac{f}{700}) \tag{1.7}$$

Un exemple de banque de filtres de Mel pour cinq filtres, en utilisant la formule de Young est donné en figure 1.9.

Les fréquences sont filtrées avec chaque filtre de Mel, en calculant le logarithme de la puissance spectrale pour chacun des filtres, ce qui donne ainsi un coefficient par filtre (voir équation 1.8).

$$E_i = \log(\sum_{k=1}^{K} w_i(k) * P_i(k)) \tag{1.8}$$

E_i représente la puissance spectrale logarithmique, $w_i(k)$ le poids du filtre numéro i associé à la $k^{\text{ième}}$ fréquence en Hertz.

Une transformée en cosinus discrète est ensuite appliquée à l'ensemble des puissances spectrales logarithmiques (voir équation 1.9).

$$MFCC_m = \sum_{i=1}^{K} E_i \cos[m(i - 0.5)\pi/K] \tag{1.9}$$

$MFCC_m$ représente le $m^{\text{ième}}$ coefficient cepstral. Seuls les premiers coefficients sont gé-

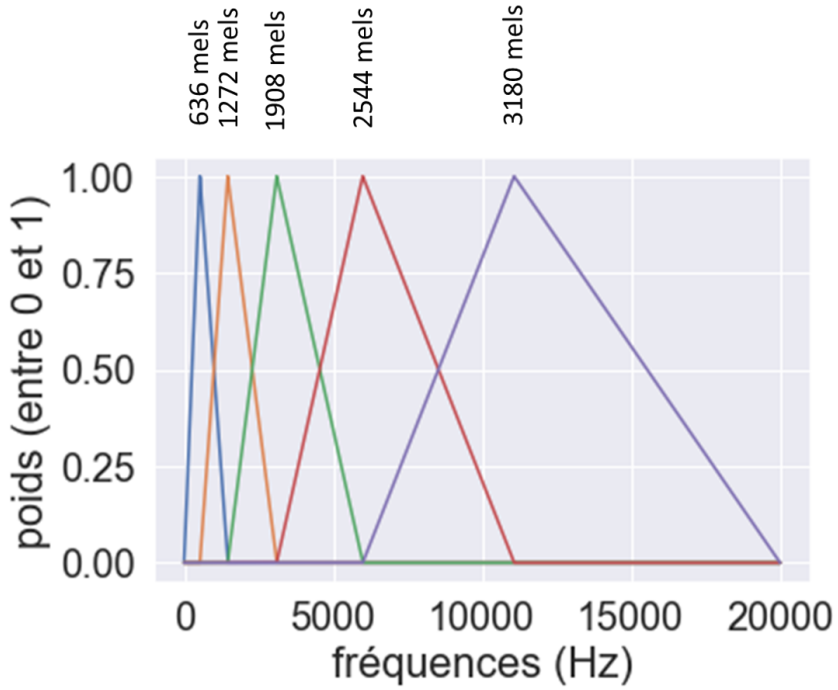

Figure 1.9 – Représentation d'une banque de filtres de Mel pour cinq filtres. Chaque couleur représente un filtre différent. Les valeurs en mels de chacune des fréquences centrales des filtres sont représentées dans la partie supérieure du graphique. Pour cinq filtres construits avec la formule de Young, les fréquences centrales sont toutes séparées de 636 mels les unes par rapport aux autres.

néralement gardés à l'issue du calcul. Une soustraction cepstrale peut être effectuée sur les MFCC afin de déconvoluer le signal du bruit lié à la source de l'enregistrement. Pour effectuer cette opération, il suffit de retirer à chaque coefficient cepstral la moyenne des coefficients cepstraux.

Le contraste spectral

Le contraste specral est introduit par Jiang *et al.* (2002). Après avoir séparé le spectre du signal en plusieurs sous-bandes représentant chacune une octave, le contraste spectral permet d'estimer l'intensité des pics spectraux et des vallées, et leurs différences dans chacune des sous-bandes. L'intensité des pics spectraux est estimée en faisant la moyenne dans le voisinage autour des valeurs de maximum et de minimum, au lieu de prendre la valeur exacte du minimum et du maximum. De ce fait, un facteur α est introduit, facteur servant à définir le voisinage autour du maximum et du minimum.

$(x_{k,1}, x_{k,2}, ..., x_{k,N})$ désigne le vecteur issu de la FFT pour la k^{ieme} bande. Tout d'abord, les valeurs de ce vecteur sont rangées par ordre décroissant, ce qui donne le vecteur $(x'_{k,1}, x'_{k,2}, ..., x'_{k,N})$. Il est alors possible d'estimer la valeur des pics (voir équation 1.10) et la valeur des vallées (voir équation 1.11) pour chaque bande k. Le contraste spectral

CS_k de chaque sous bande k peut alors être calculée via l'équation 1.12.

$$Pic_k = \log_{10}(\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x'_{k,i}) \tag{1.10}$$

$$Vallee_k = \log_{10}(\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x'_{k,N-i+1}) \tag{1.11}$$

$$CS_k = Pic_k - Vallee_k \tag{1.12}$$

1.5.4 Modèles de classification supervisée en classes multiples

De nombreux modèles de *Machine Learning* peuvent être utilisés pour des problèmes de classification en classes multiples. Beaucoup de ces méthodes ont été abordées dans les travaux d'Aly (2005).

Seuls quelques modèles ayant particulièrement prouvé leur performance pour de la classification d'évènements audio seront présentés ici. Les livres rédigés par Géron (2019), Camastra et Vinciarelli (2008), et Raschka et Mirjalili (2017) permettent d'obtenir de plus amples informations quant au fonctionnement de ces modèles, et de bien d'autres modèles qui ne seront pas abordés dans ce travail. L'algorithme des plus proches voisins, bien qu'utilisé dans de nombreux travaux (comme ceux de Azarloo et Farokhi (2012)), ne sera pas exploité dans cette étude. En effet, cette méthode est réputée pour être très coûteuse en calcul dans la phase de test (Raschka et Mirjalili, 2017), phase qui est justement cruciale pour une éventuelle application en temps réel. De même, les algorithmes de *Support Vector Machine* (SVM), bien que très utilisés (par exemple dans les travaux de Wei *et al.* (2020)), seront exclus de cette étude, car ils ne sont pas adaptés aux bases de données contenant des centaines de milliers d'observations Géron (2019).

Le perceptron multicouche

Le perceptron multicouche a été utilisé dans de multiples travaux de classification audio, comme ceux de Myint et Ni (2020), Medina *et al.* (2020), ou encore Mandal *et al.* (2020). Myint a par exemple obtenu une précision variant entre 98% et 100% en voulant classifier les évènements musicaux avec et sans voix, et Mandal entre 69% et 73% pour une classification des émotions dans la musique.

— ***principe général***

Le perceptron multicouche est un type de réseau de neurones. Chaque unité logique du réseau, c'est-à-dire chaque neurone, calcule la somme pondérée de toutes les entrées (et du

biais qui est une variable indépendante des neurones précédents), et applique au résultat une fonction d'activation (voir figure 1.10). La fonction sigmoïde ($act(z) = \frac{1}{1+e^{-z}}$) est par exemple une des fonctions d'activation les plus utilisées (voir figure 1.11).

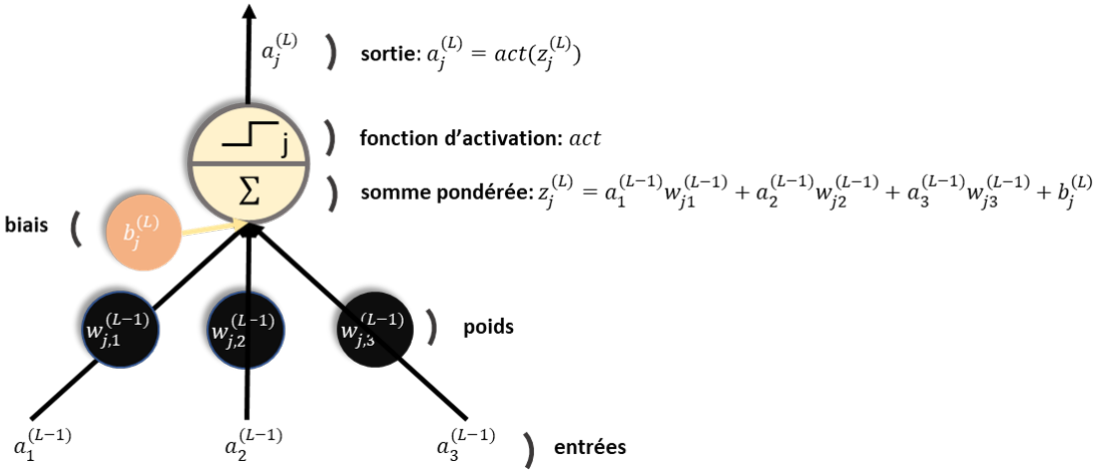

Figure 1.10 – Unité logique de la couche intermédiaire d'un réseau de neurones. Est représenté ici le $j^{\text{ième}}$ neurone de la couche L. Ce neurone prend en entrée les résultats pondérés des calculs issus des neurones de la couche précédente, et renvoie une activation.

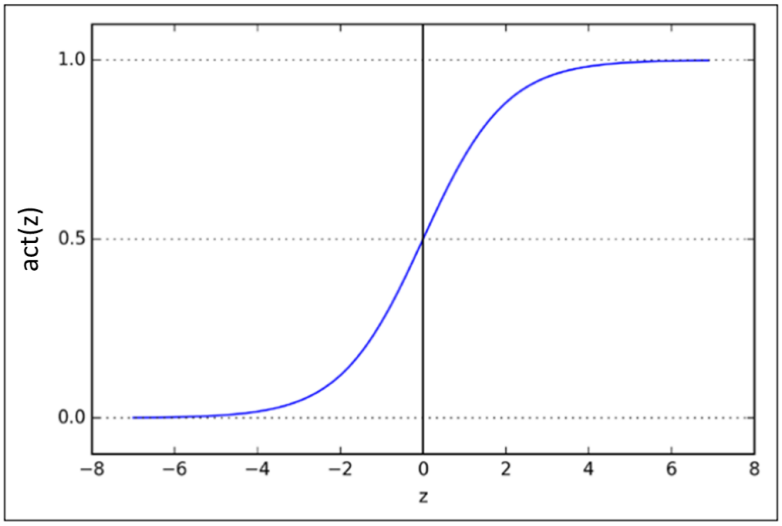

Figure 1.11 – Fonction d'activation sigmoïde (inspiré de Raschka et Mirjalili (2017)).

Pour créer un réseau profond, les unités logiques sont connectées les unes aux autres et organisées en couches, avec une couche d'entrée, une couche de sortie, et une ou plusieurs couches intermédiaires (également nommées couches cachées). Ce réseau est dès lors nommé *Deep Neural Network* (DNN) (voir figure 1.12). La couche d'entrée est passive, c'est-à-dire qu'elle se contente de faire entrer les données dans le réseau (en attribuant éventuellement un poids à chaque entrée). Dans le cas d'une classification en classes

multiples, la couche de sortie possède des neurones un peu différents de ceux de la couche intermédiaire, car la fonction d'activation *softmax* a besoin de prendre tous les neurones de la couche en entrée. Lorsqu'il n'y a pas de couche intermédiaire entre les couches d'entrées et de sorties, on parle de perceptron.

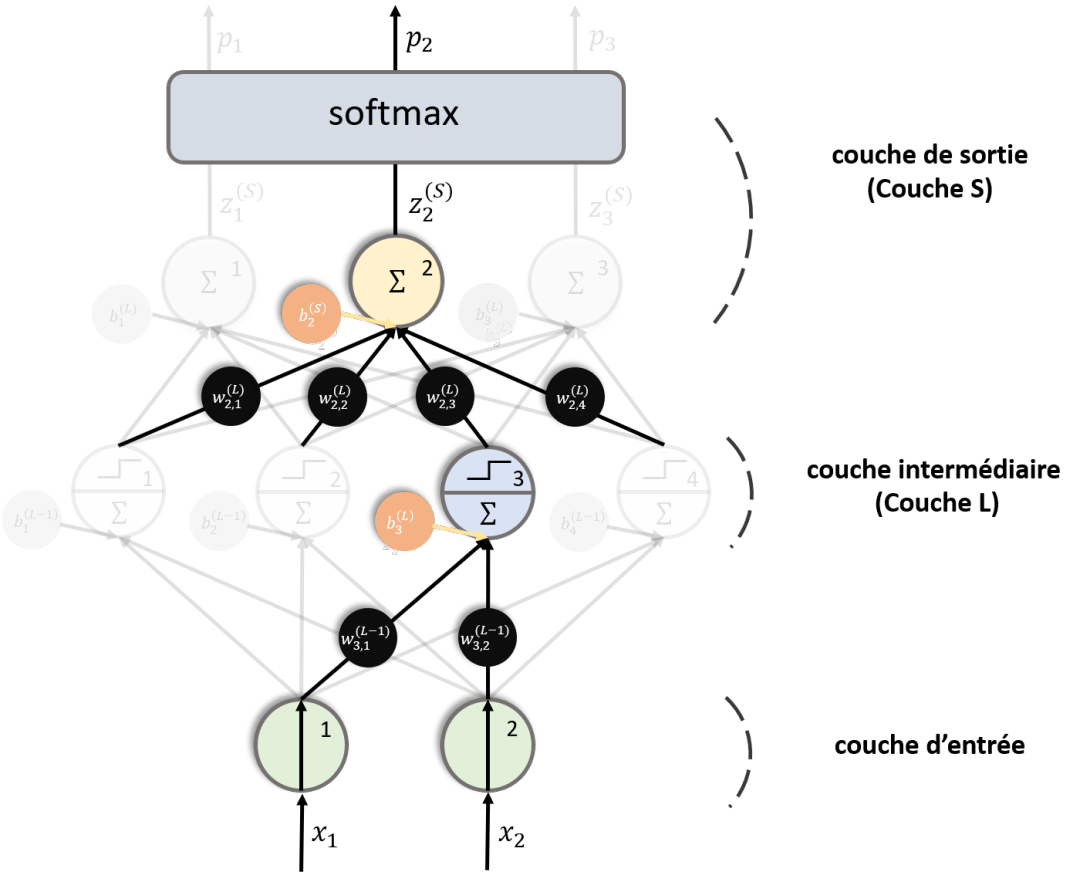

Figure 1.12 – Réseau de neurones contenant une unique couche intermédiaire composée de quatre neurones. Pour améliorer la lisibilité, seuls les poids liés au troisième neurone de la couche intermédiaire, et au deuxième neurone de la couche de sortie ont été représentés.

— ***prédiction de la classe d'une observation***

Chaque neurone de la couche de sortie sera associé à une classe, et le vecteur de probabilités $p = (p_1, p_2, ..., p_{n_S})$ permettra d'émettre une prédiction sur l'appartenance de l'observation étudiée à cette classe. Par exemple, dans le cas d'une classification en trois classes comme présentée dans la figure 1.12, on pourrait associer les probabilités p_1, p_2 et p_3 aux probabilités d'appartenance aux classes voix claire *death growl* et *hardcore scream* (suivant la classification établie en section 2.1). Une observation, lorsqu'elle entre dans le réseau de neurones, pourrait par exemple à l'issue des calculs, obtenir un vecteur de probabilités $p = (0.1, 0.7, 0.2)$, et serait alors classée dans la catégorie *death growl*.

La couche de sortie possède des neurones un peu différents de ceux de la couche intermédiaire. En effet, dans le cas où les neurones de sortie seraient du même type que

le neurone présenté en figure 1.10, la somme des sorties $\sum_{b=1}^{n_S} a_b^{(S)}$ ne serait pas forcément égale à 1. Par exemple $a^{(S)}$ pourrait être de la forme $(0.5, 0, 6, 0.4)$. Pour pallier ce problème, il est nécessaire d'employer une couche de sortie de type *softmax*. Les $z_n^{(S)}$ de chaque neurone n de la couche de sortie sont calculés exactement de la même façon que dans l'équation 1.14, mais cette fois-ci la fonction d'activation est remplacée par la fonction *softmax*, qui permet de renvoyer une probabilité p_n d'appartenance à la classe n (voir équation 1.13).

$$p_n = \frac{exp(z_n^{(S)})}{\sum_{b=1}^{n_S} exp(z_b^{(S)})} \tag{1.13}$$

Au sein de la couche intermédiaire L contenant n_L neurones, $a^{(L)} = (a_1^{(L)}, a_2^{(L)}, ..., a_{n_L}^{(L)})$ désignera le vecteur de taille n_L représentant l'ensemble des activations $a_j^{(L)}$ pour chaque sortie de neurone j. En suivant la même logique de notation, k désignera les indices des neurones d'activation $a_k^{(L-1)}$ de la couche précédente $L-1$ et n_{L-1} le nombre de neurones de cette couche. Dans le cas présenté en figure 1.12, $L-1$ est la couche d'entrée, car il n'y a qu'une seule couche intermédiaire.

Au niveau de la couche d'entrée, le vecteur $x = (x_1, x_2, ..., x_d)$ de taille d, est le vecteur contenant les descripteurs de l'observation étudiée. Dans l'exemple donné en figure 1.12, x_1 et x_2 pourraient par exemple représenter les deux premiers MFCC d'une trame (voir section 1.5.3). La couche d'entrée possède ainsi autant de neurones qu'il y a de descripteurs à étudier pour une trame, et fait rentrer chaque descripteur dans le réseau.

Pour créer un vecteur de prédiction p à partir du vecteur x, il convient d'effectuer les calculs suivants. Pour chaque neurone de la couche intermédiaire, la somme pondérée $z_j^{(L)}$ de toutes les sorties de la couche précédente est calculée, en prenant en compte le biais $b_j^{(L)}$ propre à chaque neurone (voir équation 1.14). $w_{jk}^{(L)}$ est le poids associé à la liaison entre le $k^{\text{ième}}$ neurone de la couche $L-1$ et le $j^{\text{ième}}$ neurone de la couche L. $a_j^{(L-1)}$ représente l'activation du $j^{\text{ième}}$ neurone de la couche $L-1$ (voir figure 1.10). Dans le cas où $L-1$ est la couche d'entrée, $a_j^{(L-1)}$ est en fait x_j. Les $z_n^{(S)}$ de la couche de sortie sont calculés en utilisant la même équation 1.14, et, enfin, le vecteur de sortie p est calculé grâce à l'équation 1.13.

$$z_j^{(L)} = (\sum_{k=1}^{n_{L-1}} w_{jk}^{(L)} a_j^{(L-1)}) + b_j^{(L)} \tag{1.14}$$

$$a_j^{(L)} = act(z_j^{(L)}) \tag{1.15}$$

— ***entraînement du modèle***

L'entraînement de ce modèle s'effectue par une méthode de rétropropagation. L'ensemble des données d'entraînement est regroupé en mini-batch de données, c'est-à-dire en petit groupement de données dont l'ordre a été randomisé. Par exemple, un mini-batch pourrait être constitué d'un ensemble de 64 observations. Pour chaque observation issue du mini-batch, l'algorithme de rétropropagation fait d'abord une prédiction (passe avant), mesure l'erreur qui en découle par rapport à la classe attendue de l'observation, puis parcourt

chaque couche en sens inverse pour mesurer la contribution de chaque neurone à l'erreur calculée (passe inverse), et enfin modifie légèrement les poids et biais de chaque neurone pour réduire cette erreur (étape de descente de gradient). Les étapes détaillées de calcul issues de l'entraînement du réseau de neurones sont présentées ci-dessous.

Tout d'abord, l'erreur entre les valeurs attendues et évaluées pour une observation m d'un mini-batch est estimée grâce à une fonction coût (comme par exemple l'erreur quadratique moyenne présentée en équation 1.17). On notera $Cost$ la fonction coût choisie, et C_m l'évaluation de la fonction coût pour le $m^{\text{ième}}$ fichier du mini-batch étudié (voir équation 1.16). $y = (y_1, y_2, ..., y_{n_S})$ est le vecteur de taille n_S, représentant les classes réelles d'une observation.

$$C_m = Cost(a^{(S)}, y) \tag{1.16}$$

$$C_m = \sum_{j=1}^{n_S} (a_j^{(S)} - y_j)^2 \tag{1.17}$$

Pour simplifier l'approche, les calculs suivants concernent les neurones de la couche L. Les mêmes calculs peuvent être effectués vis-à-vis des neurones de la couche S ou de la couche $L-1$. Au sein de la couche L, les dérivées partielles de la fonction coût du $m^{\text{ième}}$ fichier du mini-batch par chaque poids (équation 1.18), chaque biais (équation 1.19), et chaque activation issue de la couche précédente (équation 1.20) peuvent être calculées. Dans le cas où $L-1$ est une couche intermédiaire, la dérivée partielle de la fonction coût par rapport à chaque activation de la couche $L-1$ permettra ainsi d'y effectuer les mêmes étapes de calcul, c'est-à-dire de calculer la dérivée partielle de chacun des poids et de chacun des biais de la couche $L-1$. Si la couche $L-1$ est la couche d'entrée, il ne sera bien sûr pas nécessaire de calculer la dérivée partielle de la fonction coût par rapport à l'activation de la couche $L-1$ (puisque les activations de la couche d'entrée sont issues du vecteur x qui est une constante).

$$\frac{\partial C_m}{\partial w_{jk}^{(L)}} = \frac{\partial z_j^{(L)}}{\partial w_{jk}^{(L)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial C_m^{(L)}}{\partial a_j^{(L)}} \tag{1.18}$$

$$\frac{\partial C_m}{\partial b_j^{(L)}} = \frac{\partial z_j^{(L)}}{\partial b_j^{(L)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial C_m^{(L)}}{\partial a_j^{(L)}} \tag{1.19}$$

$$\frac{\partial C_m}{\partial a_k^{(L-1)}} = \sum_{k=1}^{n_S} \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}} \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \frac{\partial C_m^{(L)}}{\partial a_j^{(L)}} \tag{1.20}$$

La fonction coût totale est définie comme étant la moyenne des fonctions coûts de chaque fichier m du mini-batch. Pour entraîner le modèle, il faut minimiser cette fonction coût totale, c'est-à-dire trouver des valeurs de poids et de biais pour chaque neurone de chaque couche de chaque fichier m permettant d'avoir le minimum d'erreur entre les valeurs prédites issues du vecteur p et les valeurs attendues représentées par le vecteur y. Les dérivées partielles de la fonction coût totale, par rapport à chaque poids, chaque biais et chaque activation issue de la couche précédente peuvent donc être calculées à partir des dérivées partielles des fonctions coût de chaque fichier m. L'équation 1.21 montre

par exemple la dérivée partielle en $w_{jk}^{(L)}$ de la fonction coût totale. n_m désigne le nombre d'observations dans un mini-batch.

$$\frac{\partial C}{\partial w_{jk}^{(L)}} = \frac{1}{n_m} \sum_{m=1}^{n_m} \frac{\partial C_m}{\partial w_{jk}^{(L)}} \tag{1.21}$$

À partir des différentes dérivées partielles de la fonction coût issues du réseau de neurones, il est possible d'appliquer un algorithme de descente de gradient stochastique pour chaque poids et chaque biais. L'algorithme de descente de gradient est ici dit « stochastique » car le mini-batch change à chaque itération. Cela évite de traiter l'intégralité des observations à chaque étape de la descente de gradient. L'équation 1.22 présente une étape $g+1$ de la descente de gradient pour redéfinir le poids $w_{jk}^{(L)}$. Dans cette équation, ϵ représente le taux d'apprentissage, fixé par l'utilisateur de l'algorithme.

$$w_{jk}^{(L)}(g+1) = w_{jk}^{(L)}(g) + \epsilon \frac{\partial C}{\partial w_{jk}^{(L)}} \tag{1.22}$$

La descente de gradient est alors répétée pour un nouveau mini-batch, jusqu'à atteindre une convergence de l'algorithme. Il existe plusieurs algorithmes permettant d'optimiser cette descente de gradient. L'algorithme ADAM, établi par Kingma et Ba (2015) est par exemple un des algorithmes d'optimisation les plus utilisés.

La forêt d'arbres décisionnels

La forêt d'arbres décisionnels a été utilisée par exemple dans les travaux de Thambi *et al.* (2014) et Zhang et LV (2015), dans lesquels une précision de 97,8% a été obtenue pour de la détection de parole.

— *principe général*

L'arbre décisionnel est composé d'un ou de plusieurs nœuds de décision, qui permettent chacun de répartir les données dans deux sous-ensembles (voir figure 1.13). À partir d'un nœud dit racine, l'arbre peut se déployer en plusieurs nœuds intermédiaires avant d'atteindre les nœuds feuilles (voir figure 1.16). Chaque nœud feuille sera intrinsèquement lié à une liste de probabilités d'appartenance à chacune des classes. Par exemple, pour une classification en trois classes, un nœud i pourrait être associé à une probabilité $p_{i1} = 0,1$ de voix claire, de $p_{i2} = 0,7$ de death growl, et de $p_{i3} = 0.2$ de hardcore scream. Lorsqu'une observation atteint ce nœud feuille, elle sera alors associée à la catégorie death growl, car il s'agit de la catégorie possédant la probabilité $p_{i2} = 0,7$ la plus élevée dans ce nœud. La profondeur maximale d'un arbre de décision peut-être définie comme le nombre maximal de nœuds possibles entre le nœud racine et un des nœuds feuilles (le nœud feuille étant comptabilisé). Par exemple, l'arbre de décision de la figure 1.14 a une profondeur maximale de 4.

La forêt d'arbres décisionnels est un ensemble de forêts de décisions, conçu en décou-

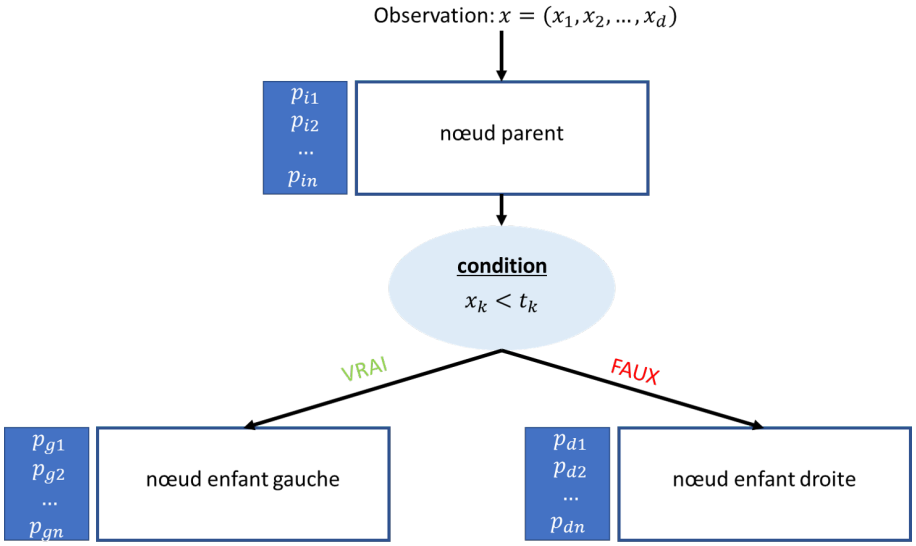

Figure 1.13 – Exemple de nœud de décision d'indice i.

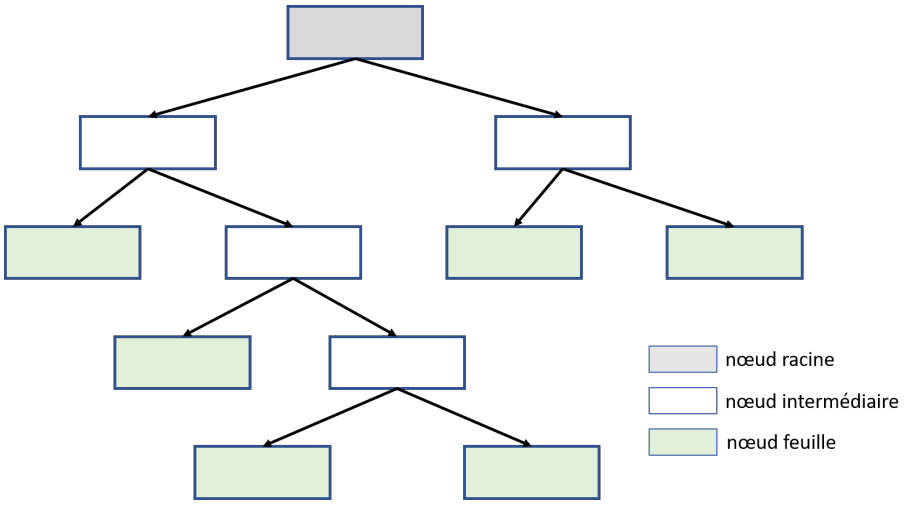

Figure 1.14 – Exemple d'un arbre de décision possédant une profondeur maximale de 4.

pant les données d'entraînement en sous-groupes de façon aléatoire, en en entraînant un arbre décisionnel différent pour chaque sous-groupe (voir figure 1.16). Si une observation des données d'entraînement peut se trouver plusieurs fois dans le même arbre décisionnel, on parlera de technique de *bagging*. Si elle ne peut être utilisée qu'une seule fois par arbre, on parlera de technique de *pasting*. Peu importe la technique utilisée, une même observation des données pourra toujours être utilisée dans plusieurs arbres décisionnels.

— ***prédiction de la classe d'une observation***

Pour évaluer la classe d'une observation, il faut faire entrer ses descripteurs dans l'arbre décisionnel. L'algorithme fait passer l'observation du nœud parent à un des deux nœuds

enfant en vérifiant si un descripteur spécifique répond ou non à la condition imposée par le seuil (voir figure 1.13). L'observation passe de nœuds en nœuds jusqu'à atteindre un nœud feuille. La classe estimée pour cette observation sera la classe représentée par l'indice k du maximum des p_{ik} de cette feuille. Dans le cas d'une forêt d'arbres décisionnels, cette évaluation doit être effectuée pour chacun des arbres. La classe prédite par la forêt sera alors celle prédite par la majorité des arbres de la forêt.

— ***entraînement du modèle***

Les arbres décisionnels sont construits lors de la phase d'entraînement du modèle, en évaluant récursivement différents descripteurs et en utilisant pour chaque nœud les descripteurs qui séparent le mieux les données. Le score de chaque nœud peut en effet être évalué à partir de l'indice de diversité de Gini (voir équation 1.23). G_i représente l'indice de diversité de Gini du $i^{\text{ième}}$ nœud, et p_{ik} est alors calculé comme étant le ratio des observations de classe numéro k parmi les données d'entraînement passant dans le $i^{\text{ième}}$ nœud.

$$G_i = 1 - \sum_{k=1}^{n} p_{ik}^2 \tag{1.23}$$

L'algorithme CART (Classification And Regression Tree) permet de construire cet arbre de décision à partir des données d'entraînement. L'algorithme commence par trouver le descripteur k et le seuil t_k permettant d'obtenir un nœud dont les feuilles sont les plus pures possibles en minimisant la fonction coût présentée en équation 1.24. La fonction coût peut être minimisée grâce à un algorithme de descente de gradient (dont le principe algorithmique est identique à celui présenté en équation 1.22). L'indice de Gini et l'algorithme CART ont été introduits pour la première fois dans les travaux de Breiman *et al.* (1984).

$$J(k, t_k) = \frac{m_{gauche}}{m} G_{gauche} + \frac{m_{droite}}{m} G_{droite} \tag{1.24}$$

$m_{gauche/droite}$ représente le nombre d'occurrences dans le nœud enfant de gauche ou de droite, et $G_{gauche/droite}$ l'indice de diversité du nœud enfant de gauche ou de droite.

Un exemple de calcul de cette fonction coût est donné en figure 1.15.

Une fois que le premier nœud décisionnel est conçu, la même opération est effectuée pour chacun des nœuds enfant afin de séparer de nouveau les données en deux pour chaque nœud. Si l'algorithme ne parvient pas à trouver des nœuds enfants qui réduiraient l'indice de diversité d'un nœud parent (si pour tout paramètre k et seuil t_k $G_{enfant} > G_{parent}$), alors le nœud parent en question devient un nœud feuille. L'algorithme s'arrête lorsqu'il a transformé tous les nœuds en cours de calcul en nœuds feuille, ou lorsqu'il atteint une ou plusieurs conditions spécifiques (par exemple s'il atteint la profondeur maximale spécifiée par l'utilisateur de l'algorithme).

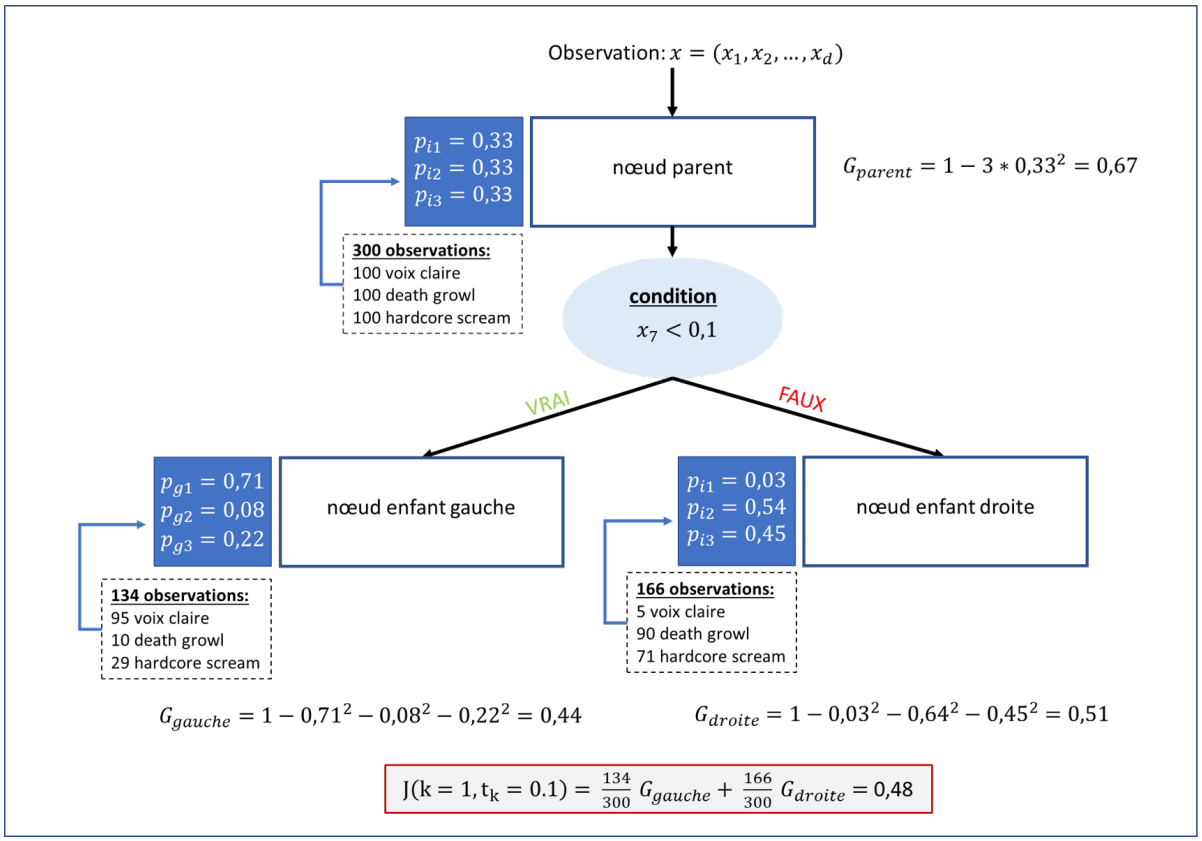

Figure 1.15 – Exemple de calcul du $J(k, t_k)$ avec k=7 et $t_k = 0,1$.

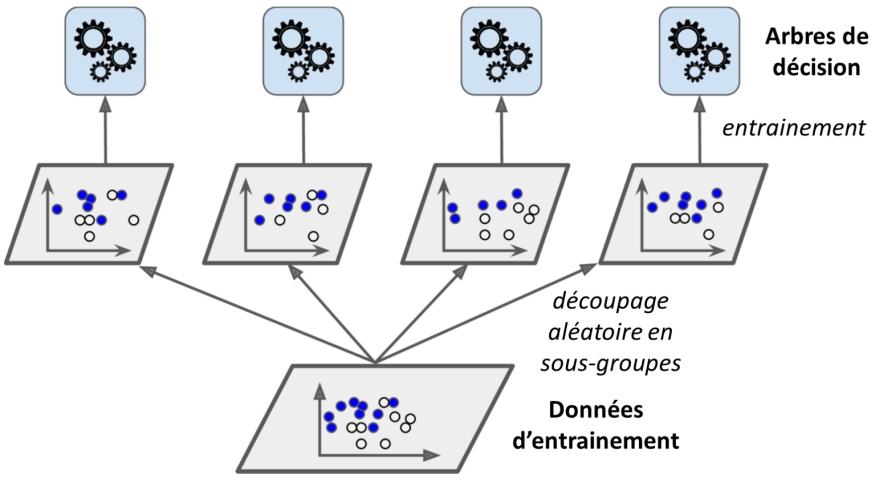

Figure 1.16 – Fonctionnement d'un arbre de décision.

La classification naïve bayésienne avec loi gaussienne

La classification naïve bayésienne est une méthode très simple pouvant donner des résultats très satisfaisants. Iheme et Ozan (2019) a obtenu un maximum de précision de 87% pour une classification en trois classes pour de la parole, de la musique, et du silence. Cette méthode, bien que très efficace, est cependant très souvent moins performante que les algorithmes plus récents comme les algorithmes de *Support Vector Machine* ou de forêt d'arbres décisionnels (voir étude de Caruana et Niculescu-Mizil (2006)).

— ***principe général***

La classification naïve bayésienne est basée sur la loi de Bayes présentée en équation 1.25. Ici, $x = (x_1, x_2, ..., x_d)$ représente le vecteur contenant d valeurs de chaque descripteur issu d'une observation appartenant à la base de données de test, et y_j représente la classe j.

$$P(y_j|x) = \frac{P(x|y_j)P(y_j)}{P(x)} \tag{1.25}$$

— ***prédiction de la classe d'une observation***

Pour une observation x de la base de données de test, la probabilité $P(y_j|x)$, également appelée vraisemblance, est calculée pour chacune des classes j. La classe possédant la vraisemblance la plus élevée sera la classe prédite pour l'observation x.

— ***entraînement du modèle***

Contrairement aux méthodes de perceptron multicouche et de forêt d'arbres décisionnels, ici l'entraînement du modèle s'effectue par un simple calcul de la variance et de l'écart type des données d'entraînement, et ce pour chaque classe j et chaque paramètre k.

En effet, si l'on suppose que les paramètres sont indépendants les uns par rapport aux autres, on peut représenter $P(x|y_j)$ sous la forme d'un produit de probabilités conditionnelles de chaque paramètre (voir équation 1.26).

$$P(x|y_j) = \prod_{k=1}^{d} P(x_k|y_j) \tag{1.26}$$

En supposant une distribution gaussienne pour les lois de probabilité des paramètres, on peut évaluer $P(x_k|y_j)$ en calculant l'espérance et l'écart type des données d'entraînement pour chaque classe j et pour chaque paramètre k. $\sigma_{k,j}$ désigne cet écart type, et $\mu_{k,j}$ cette espérance. La fonction gaussienne est alors utilisée pour calculer $P(x_k|y_j)$ (voir équation 1.27).

$$P(x_k|y_j) = \frac{1}{\sqrt{2\pi\sigma_{k,j}}} \exp^{-\frac{(x_k-\mu_{k,j})^2}{2\sigma_{k,j}^2}} \tag{1.27}$$

$P(y_j)$ est la proportion d'observations de classe j au sein des données d'entraînement, N désignant le nombre total d'observations au sein des données d'entraînement (voir équation 1.28).

$$P(y_j) = \frac{N_j}{N} \tag{1.28}$$

Le terme $P(x)$ n'a pas besoin d'être calculé, car il agit simplement en facteur d'échelle dans l'équation 1.25. Plutôt que de comparer les vraisemblances, on compara donc les termes $P(x|y_j)P(y_j)$ entre chaque classe j (voir inéquation 1.29)

$$P(y_j|x) > P(y_i|x) \Leftrightarrow \frac{P(x|y_j)P(y_j)}{P(x)} > \frac{P(x|y_i)P(y_i)}{P(x)} \Leftrightarrow P(x|y_j)P(y_j) > P(x|y_i)P(y_i) \tag{1.29}$$

1.6 Conclusion pour l'état de l'art

Les taxonomies existantes de techniques de distorsion vocale les plus extrêmes sont très éparses. Cependant, il est possible de distinguer des techniques différentes à partir de l'histoire de ces techniques vocales, en les rattachant à un genre musical. Un des objectifs de ce travail sera donc d'établir une nouvelle classification de ces techniques. Les précédentes recherches en classification de techniques vocales extrêmes ont montré que ces techniques peuvent être discriminées par des algorithmes de *Machine Learning*. Les MFCC semblent être les descripteurs les plus adaptés pour l'analyse de ces techniques vocales. Plusieurs modèles de *Machine Learning* ont obtenu de très bonnes performances pour de la classification supervisée en classes multiples appliquée à des signaux vocaux, il conviendra donc de comparer leurs performances.

Chapitre 2

Méthodes

2.1 Taxonomie choisie pour la création de la base de données

. En considérant les travaux des différents chercheurs ayant travaillé sur les techniques de distorsion vocale les plus extrêmes utilisées dans le *metal*, une nouvelle taxonomie composée du *black shriek*, du *death growl* du *hardcore scream* et du *grind inhale* est élaborée. L'absence d'une taxonomie universellement utilisée rend les dénominations actuelles parfois peu intuitives pour les chanteurs. Ici chaque technique vocale est donc associée à un sous-genre musical du *heavy metal*, rattachant de ce fait implicitement cette nouvelle taxonomie à des exemples concrets.

Ces techniques vocales sont également découpées en trois registres différents, nommés *high*, *mid* et *low*. Le *high* correspond à une hauteur proche de la voix de tête du chanteur, le *mid* correspond à une hauteur proche de sa voix parlée, et le *low* à une hauteur en deçà de sa voix parlée. Il est à noter qu'ici le *black shriek* n'a pas de registre *low*, et que le *death growl* n'a pas de registre *high*. Cela fait suite aux conclusions de Hainaut (2020), qui constate que les voix du *black metal* sont majoritairement plus aiguës que celles du *death metal*. Ce constat a été confirmé par des chanteurs de *metal* questionnés avant d'effectuer les enregistrements, qui ont pu préciser ce propos en affirmant qu'une technique *black shriek* dans le registre *low* serait un *death growl* et qu'une technique *death growl* dans le registre *high* serait un *black shriek*. Le *grind inhale*, correspondant au genre *grindcore*, n'est présenté que dans un seul registre. Cette technique peut être effectuée dans les registres *high*, *mid* et *low*, mais suite aux remarques de plusieurs chanteurs attestant que cette technique est dangereuse pour la santé des cordes vocales, les enregistrements n'ont été effectués ici que pour un seul registre. Le *grind inhale* est d'ailleurs de moins en moins représenté dans le *heavy metal*.

Plusieurs raisons ont mené au choix de découper les techniques en registres et de ne pas imposer une note particulière aux chanteurs. La première est que la pratique de techniques de distorsion extrême est très peu académique et se concentre plus sur l'agressivité

que sur l'intelligibilité et la justesse des notes, ce qui fait que les chanteurs sont souvent peu habitués à ce que leur soit demandée la production d'une note spécifique lorsqu'ils emploient des techniques vocales saturées. La seconde est que pour l'application de méthodes de *Machine Learning*, il est plus judicieux d'avoir des hauteurs de voix différentes entre les chanteurs pour diversifier la base de données. Enfin, la hauteur tonale devenant non identifiable pour les techniques les plus extrêmes (Nieto, 2008; Smialek *et al.*, 2012), demander la production d'une note particulière peut s'avérer peu pertinent sur le plan acoustique.

2.1.1 Catégories de distorsion vocale

Le *black shriek*

Le *black shriek* est très utilisé dans le *black metal*, mais également dans le *death metal*. Il est généralement plus aigu que le *death growl*, la différence se ressentant particulièrement dans la prononciation des voyelles, avec un timbre qui laisse entendre un son vocalique intermédiaire entre le le [a] et le [ɛ] (comme dans « mère ») (Hainaut, 2020). Du fait de sa nature aiguë, il est utilisé dans les tonalités *high* et *mid*, mais pas dans la tonalité *low*.

Cette technique vocale est par exemple représentée par les chanteurs des groupes *Marduk*, *Satyricon*, *Behemoth*, *Dimmu Borgir* ou *Darkthrone*. Le *black shriek* dans le registre *high* peut être entendu dans le morceau *Nemesis* de *Cradle Of Filth* (issu de l'album Nymphetamine (2004)), et dans le registre *mid* peut être entendu dans le morceau *Goat Of Mendes* de *Impaled Nazarene* (issu de l'album Eight Headed Serpent (2021)).

Le *death growl*

Le *death growl* est quant à lui majoritairement utilisé dans le *death metal*, mais également dans le *thrash metal* ou le *black metal*. Concernant la prononciation des voyelles, le *death growl* laisse entendre un son vocalique intermédiaire entre le [a] et le [ɔ] (comme dans « grotte ») (Hainaut, 2020). Le *death growl* est généralement grave, et se retrouve dans les tonalités *mid* et *low*.

Cette technique vocale est par exemple représentée par les chanteurs des groupes *Cannibal Corpse*, *Cattle Decapitation*, *Deicide* ou *Gorod*. Le *death growl* dans le registre *Mid* peut être entendu dans le morceau *Divine Immolation* de *Schizophrenia* (issu de l'album Recollections of the Insane (2022)), et dans le registre *low* dans le morceau *Creature* de *Baest* (issu de l'album Creature (2022)).

Le *hardcore scream*

Le *hardcore scream*, souvent utilisé dans le *hardcore* et le *grindcore*, désigne une voix criée plus ou moins bruitée. Cette technique est également utilisée dans le registre

high dans beaucoup de morceaux de *black metal*. La distorsion induite par cette technique produit des harmoniques moins graves que pour les techniques de *death growl* et de *black shriek*. *A priori*, ce type de vocalise est utilisé principalement dans les tonalités *high et mid* mais peut également être trouvé de façon plus rare dans le registre *low*.

Cette technique vocale est par exemple représentée par les chanteurs des groupes *Slipknot*, *The Dillinger Escape Plan*, *Stick To Your Guns*, *Fall In Archaea*, *Terror*, *Get The Shot* ou *Silverstein*. Le *hardcore scream* dans le registre *high* peut être entendu dans le morceau *Cold Hearted* de *Get The Shot* (issu de l'album No Peace in Hell (2014)), dans le registre *mid* dans le morceau *Against Them All* de *Stick To Your Guns* (issu de l'album Diamond (2012)), et dans le registre *low* dans *You Will Never Be One Of Us* de *NAILS* (issu de l'album You Will Never Be One Of Us (2016)). Dans le genre *black metal*, on peut entendre un *hardcore scream* dans le registre *high* dans le morceau *Nuages* de *Wiegedood* (issu de l'album There's Always Blood At The End Of The Road (2022)).

Le *grind inhale*

Le *grind inhale* est un cri produit en aspirant l'air, contrairement aux autres techniques de cette liste qui sont produites en expirant. Elle peut être utilisée dans beaucoup de genres de *metal* différents, est assez dominante dans le *grindcore*, mais reste une technique globalement plus secondaire que celles citées précédemment.

Cette technique vocale est par exemple représentée par les chanteurs des groupes Annotations Of An Autopsy, Walking The Cadaver ou Archspire. Le *grind inhale* peut être entendu dans le morceau *Welcome To Sludge City* du groupe *Annotations Of An Autopsy* (issu de l'album Welcome To Sludge City (2007)).

2.1.2 Effets de distorsion vocale

Seront également étudiés plusieurs effets pouvant être utilisés pour ponctuer les différentes vocalises. Ces effets ne peuvent être considérés comme des catégories en tant que telles, car il est très rare qu'un chanteur n'utilise que ces effets lors d'un morceau. À terme, et après analyse, chacun de ces effets sera rattaché à une des catégories citée précédemment.

Le *pig squeal*

Effectuer un *pig squeal* consiste à imiter un cri de cochon. Ce cri est généralement aigu, et est majoritairement produit en aspirant l'air plutôt qu'en l'expirant (comme pour le *grind inhale*). Dans *The Man Who Built A God* de *Genocide Of Prescription* (issu de l'album Corporal Violence (2009)), un *pig squeal* est produit par le chanteur de 1 min à 1 min 6 s.

Le *deep gutturals*

Les *deep gutturals* sont des *death growls* dans l'extrême grave de la tessiture du chanteur, qui va ainsi tenter de produire la vocalise la plus grave possible. Les paroles en deviennent inintelligibles (si appliqués à un texte écrit). Dans *The Pale Horse 2.0* de *Dead/Awake* (issu de l'album Insurrectionist (Deluxe)), un *deep gutturals* est produit par le chanteur de 2 min 53 s à 2 min 59 s.

Le *tunnel throat*

La technique de *tunnel throat* est produite en recroquevillant la langue contre le palet en produisant un son de *death growl* très grave. Le son produit est généralement légèrement plus aigu que celui des *deep gutturals*. Dans *Death Atlas* de *Cattle Decapitation* (issu de l'album Death Atlas (2019)), un *tunnel throat* est produit par le chanteur de 2 min 37 s à 2 min 46 s.

2.2 Méthodes d'enregistrement pour la création de la base de données

Afin de créer une base de données permettant d'effectuer une détection des différents types de voix saturée, des chanteurs de *metal* effectuant chacune des techniques ont été enregistrés. Une catégorie a été ajoutée à la taxonomie présentée en figure 2.1 : la catégorie voix claire, en trois registres (*low*, *mid*, *high*), la voix claire désignant le chant non saturé. 27 chanteurs et chanteuses ont été enregistrés, appartenant tous au minimum à un groupe de musique se produisant en concert. Parmi ces chanteurs, cinq sont professeurs de chant.

		Registre	Exemple
CATEGORIES	HARDCORE SCREAM	High	Cold Hearted - Get The Shot
		Mid	Against Them All - Stick To Your Guns
		Low	You Will Never Be One Of Us - Nails
	BLACK SHRIEK	High	Nemesis - Cradle Of Filth
		Mid	Goat Of Mendes - Impaled Nazarene
	DEATH GROWL	Mid	Divine Immolation - Schizophrenia
		Low	Creature - Baest
	GRIND INHALE	-	Welcome To Sludge City - Annotations Of An Autopsy
EFFETS	PIG SQUEAL	-	The Man Who Build A God - Genocide Of Prescription
	DEEP GUTTURALS	-	The Pale Horse 2.0 - Dead/Awake
	TUNNEL THROAT	-	Death Atlas - Cattle Decapitation

Figure 2.1 – Taxonomie des techniques de distorsion vocales utilisées dans le *metal* extrême.

2.2.1 Matériel et lieu choisi pour les enregistrements

Le matériel suivant a été utilisé afin d'effectuer les enregistrements :

— Un microphone SM58. Ce microphone est un des microphones les plus utilisés sur scène.

— Un casque audio fermé, pour diffuser la musique et pour le retour de la voix du chanteur s'il souhaite entendre sa voix.

— Une interface audio *Scarlett 6i6* de la marque *Focusrite*. Cette interface audio est très petite, et permet donc d'avoir un matériel mobile.

— Un ordinateur portable.

Les enregistrements n'ont pas été effectués en studio, mais dans des lieux différents pour chaque chanteur, la plupart du temps à leur domicile. Plusieurs raisons à cela :

— Le microphone SM58 étant un microphone dynamique, l'acoustique de la salle captée par le microphone lors de la production de la voix est moins prononcée que pour les microphones électrostatiques habituellement utilisés en studio (Neumann U87, TLM103 etc...).

— L'algorithme développé doit pouvoir s'adapter à des environnements différents. Si les enregistrements ne sont effectués que dans un studio insonorisé, il se pourrait qu'il ne puisse fonctionner que dans cet environnement.

— Cela a permis d'enregistrer plus de chanteurs que dans les études habituelles, car les chanteurs n'avaient ainsi pas besoin de se déplacer vers un lieu d'enregistrement commun. Le fait de pouvoir enregistrer des chanteurs provenant de villes différentes (Paris et Nantes), a également permis d'enregistrer un plus grand nombre de chanteurs.

D'autres enregistrements ont été effectués en autonomie par certains chanteurs, possédant leur propre matériel. Ils avaient à leur disposition un tutoriel vidéo ainsi que des documents explicatifs leur permettant de participer au projet. Cette méthode d'enregistrement n'a été proposée qu'en dernier recours aux chanteurs n'habitant pas la région parisienne et ne pouvant pas s'y rendre, afin de multiplier les données et de permettre à des chanteurs très expérimentés de participer à l'élaboration de cette base de données.

2.2.2 Préparation des enregistrements

Les chanteurs étaient invités à s'échauffer avant les enregistrements, et à se mettre du mieux possible dans des conditions de performance scénique. La taxonomie (voir figure 2.1) était présentée aux chanteurs, avant l'enregistrement de chaque catégorie. Il était alors proposé au chanteur d'écouter un exemple de la technique qu'il s'apprête à produire, et ce avant chaque enregistrement (voir liste des exemples en Annexe B). L'ordre d'enregistrement de chaque catégorie était décidé par le chanteur. Le chanteur pouvait ne pas effectuer les enregistrements pour certaines catégories, s'il considérait qu'il ne maîtrisait pas assez les techniques en question. Cette décision fait suite aux différentes recherches

sur la dangerosité pour les cordes vocales dans la pratique de techniques de distorsion vocale extrêmes lorsque le chanteur n'y est pas préparé (McGlashan *et al.*, 2007).

Le chanteur pouvait également choisir une distance au microphone, qui restera la même pour tous les enregistrements de toutes les catégories, décidée en fonction de l'utilisation usuelle qu'en fait le chanteur sur scène. Il est cependant à noter que dans des conditions réelles de concert, le chanteur peut être amené à changer sa distance au microphone en fonction des techniques pratiquées. Ici, le choix a été fait de garder la même distance pour un même chanteur tout au long des enregistrements et changer plutôt le niveau de pré-amplification du microphone entre chaque technique, afin de pouvoir comparer les résultats issus de plusieurs techniques pour un même chanteur. De plus, les changements de position du microphone employés par les chanteurs sur scène dépendent énormément des techniques employées dans une chanson, car si le chanteur ne pratique que des techniques vocales produisant un volume sonore ressenti équivalent, il n'aura pas besoin de changer sa position au microphone.

2.2.3 Réalisation des enregistrements

Chaque chanteur devait produire trois voyelles pendant cinq secondes : [a] comme dans « arbre », [i] comme dans « livre » et [u] comme dans « ours ». Ces trois voyelles correspondent aux voyelles les plus éloignées du triangle vocalique Kent et Vorperian (2018). Plusieurs voyelles ont ainsi été enregistrées pour cette étude suite aux conclusions de Hainaut (2020) sur les différences de prononciation notables des voyelles entre les voix issues du *black metal* et celles issues du *death metal*. Le chanteur devait garder du mieux possible la même hauteur, à la fois lors de la production d'une voyelle, mais également entre les voyelles produites. Après avoir produit chaque voyelle, il devait produire une performance d'environ 15 s avec des paroles de son choix en employant cette même technique. Les paroles devaient être identiques entre chaque catégorie.

Lors des enregistrements, une boucle musicale de 5 s pour les voyelles, et de 15 s pour la partie textuelle, comprenant batterie, basse, et guitare, était diffusée dans le casque du sujet, afin de le mettre dans des conditions de concert. Une fois les enregistrements terminés avec une catégorie, la catégorie suivante était enregistrée en suivant les mêmes étapes.

2.2.4 Évaluation des enregistrements

Après les enregistrements, le chanteur était amené à s'auto-évaluer pour chaque technique vocale, se donnant une note allant de 0 à 5 en suivant le barème suivant :

— 0 : je ne maîtrise pas assez cette technique pour en enregistrer un échantillon,

— 1 : je ne pratique jamais cette technique sur scène,

— 2 : je ne pratique presque jamais cette technique sur scène,

— 3 : je pratique cette technique occasionnellement sur scène,

— 4 : je pratique cette technique assez régulièrement sur scène,

— 5 : je pratique cette technique régulièrement sur scène.

Ce tableau d'auto-évaluation permettra éventuellement d'exclure les enregistrements avec les notes les plus médiocres, si les résultats en l'état ne sont pas concluants. La liste des informations détaillées comprenant le tableau d'auto-évaluation, le matériel utilisé, et d'autres informations pouvant s'avérer utiles ont été présentées en Annexe A.

2.3 Découpage des données, descripteurs utilisés et test de modèles de *Machine Learning*

Plusieurs descripteurs, et plusieurs types d'algorithme de *Machine Learning* sont testés, et leurs performances sont évaluées. Dans un second temps, les modèles et les descripteurs utilisés pourront être optimisés.

2.3.1 Découpage des données

Tous les fichiers audio sont tout d'abord normalisés en volume, afin que leur amplitude ait un maximum légèrement inférieur à 1 et un minimum égal à -1. Le découpage des données se fait ensuite en trames de 1024 échantillons, avec un un taux de recouvrement de 50%, soit 512 échantillons. Une fenêtre de pondération de *Hann* est appliquée à chaque trame. Chaque trame représente ainsi une donnée d'entrée pour l'extraction des descripteurs.

Lors de la phase de test, une cross-validation en N-folds est effectuée, chaque pli étant associé à un chanteur (voir section 1.5.2).

2.3.2 Extraction des descripteurs

Extraction des Mel Frequency Cepstral Coefficients (MFCC)

Les 13 premiers MFCC sont extraits de chaque trame, en fixant le nombre de filtres de Mel à 128. Le nombre de MFCC à extraire, la formule utilisée pour la conversion de Hz à mels (Slaney ou Young) et le nombre de filtres de Mel pourront éventuellement être modifiés par la suite.

Extraction des Data Adjusted Frequencies Cepstral Coefficients (DAFCC)

Inspiré des MFCC, un nouveau type de descripteur conçu lors de ce mémoire est également testé. Les fréquences centrales des filtres de Mel utilisées dans le calcul de

MFCC sont remplacées par d'autres fréquences, afin d'obtenir des sous-bandes adaptées à cette étude.

Pour calculer ces nouvelles fréquences centrales, la moyenne des spectres en puissance des trames de chaque catégorie est calculée. Le spectre en puissance est obtenu via l'équation 1.3. La moyenne des puissances associées à chaque fréquence pour chacune des catégories est calculée en équation 2.1. $P(k)^{(C_j)}$ représente la puissance moyenne de la $k^{\text{ième}}$ fréquence de la classe j, avec N_j le nombre d'observations dans la classe j et $P_i(k)$ la puissance de la $k^{\text{ième}}$ fréquence de la $i^{\text{ième}}$ observation de la classe j.

$$P(k)^{(C_j)} = \frac{1}{N_j} \sum_{i=1}^{N_j} P_i(k)^{(C_j)} \tag{2.1}$$

L'ensemble des $P(k)^{(C_j)}$ associés à la catégorie j constitue ainsi le spectre de puissance moyenne de cette catégorie. Afin de mesurer l'écart de puissance moyenne entre toutes les catégories pour chaque fréquence, l'écart-type de chacune des fréquences des spectres de puissance moyenne de chaque catégorie est calculé (voir équation 2.2). σ représente ici la fonction écart-type, $\sigma(k)$ l'écart-type entre les puissances moyennes des catégories pour la $k^{\text{ième}}$ fréquence. N_c représente le nombre total de catégories. L'ensemble des $\sigma(k)$ constitue un spectre des écarts inter-catégories.

$$\sigma(k) = \sigma(P(k)^{(C_1)}, P(k)^{(C_2)}, ..., P(k)^{(C_{N_c})}) \tag{2.2}$$

Si une fréquence possède une grande amplitude dans ce spectre d'écarts inter-catégories, cela signifie ainsi que les valeurs de puissances associées à cette fréquence sont particulièrement différentes entre les catégories. A l'inverse, si une fréquence possède une faible amplitude pour ce spectre, cela signifie que les valeurs de puissances associées à cette fréquence sont très proches entre les différentes catégories. Dans l'exemple simple présenté en figure 2.2, entre 4 kHz et 9 kHz le spectre d'écarts inter-catégories a une amplitude nulle, puisque la puissance moyenne de chacune des trois classes pour ces fréquences est identique. A l'inverse, entre 9 kHz et 13 kHz, le spectre d'écarts inter-catégories a une amplitude non-nulle puisque les puissances moyennes associées à chacune des fréquences de chacune des catégories présentent des différences.

Il est alors possible de calculer le centroïde spectral du spectre d'écarts inter-catégories. Pour un spectre d'énergie, le centroïde spectral correspond à la fréquence qui permet de séparer le spectre en deux sous-bandes d'énergie égale (voir équation 2.3).

$$f_{centroide} = \frac{\sum_{k=1}^{K} f(k)A(k)}{\sum_{k=1}^{K} A(k)} \tag{2.3}$$

où $f_{centroide}$ désigne le centroide spectral, $f(k)$ la fréquence d'indice k et $A(k)$ l'amplitude spectrale qui lui est associée, tandis que K est le nombre de fréquences dans le spectre

étudié. En appliquant le calcul du centroïde spectral au spectre d'écarts inter-catégories, cette équation permet alors de calculer un centroïde spectral d'écarts inter-catégories, qui correspond ainsi à la fréquence qui partage le spectre d'écarts inter-catégories en deux sous-bandes pour lesquelles les cumuls d'écarts inter-catégories sont égaux. Autrement dit, le spectre d'écarts inter-catégories permet de séparer le spectre en deux sous-bandes qui permettent aussi bien l'une que l'autre de différencier les catégories.

De façon récursive, le centroïde spectral d'écarts inter-catégories de chaque sous bande peut-être calculé, ainsi que le centroïde spectral d'écarts inter-catégories de chaque sous-sous bande, etc. Il est possible d'obtenir de cette manière n valeurs de fréquences, n étant une puissance de deux. Ces fréquences permettent de séparer le spectre d'écarts inter-catégories en n+1 sous-bandes, permettant chacune de différencier autant les catégories. Les centroides spectraux de chacune des n+1 bandes sont de nouveau calculés, et ce sont ces centroides pour chaque sous bande qui seront utilisés comme fréquences centrales des différents filtres (voir figure 2.3). Dans les faits, cela revient à calculer $2n + 1$ valeurs de fréquences par la méthode de calcul récursive des centroïdes, et à ne garder que les fréquences impaires. Un vecteur $f_{mid} = (f_{mid_1}, f_{mid_2}, ..., f_{mid_{n+1}})$ est ainsi obtenu.

Les mêmes méthodes de calcul que celles employées pour les MFCC sont dès lors utilisées, en remplaçant simplement les fréquences centrales des filtres de Mel par ces nouvelles fréquences. Comme pour les MFCC, il est alors possible de choisir un certain nombre de ces coefficients comme descripteurs du modèle de *Machine Learning*. Ces nouveaux coefficients sont nommées les DAFCC (Data Adjusted Frequency Cepstral Coefficients).

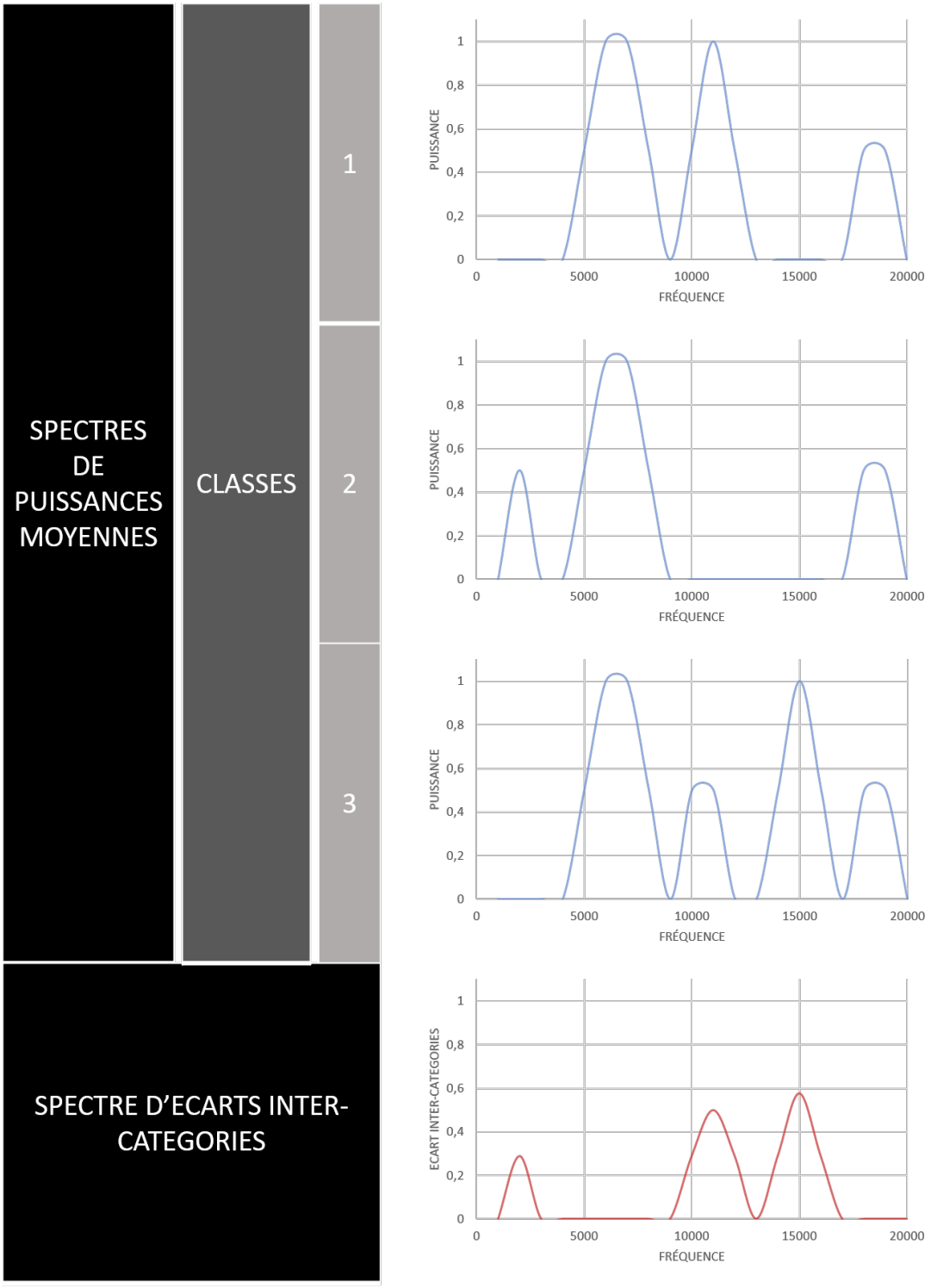

Figure 2.2 – Spectres de puissances moyennes et spectres d'écarts inter-catégories.

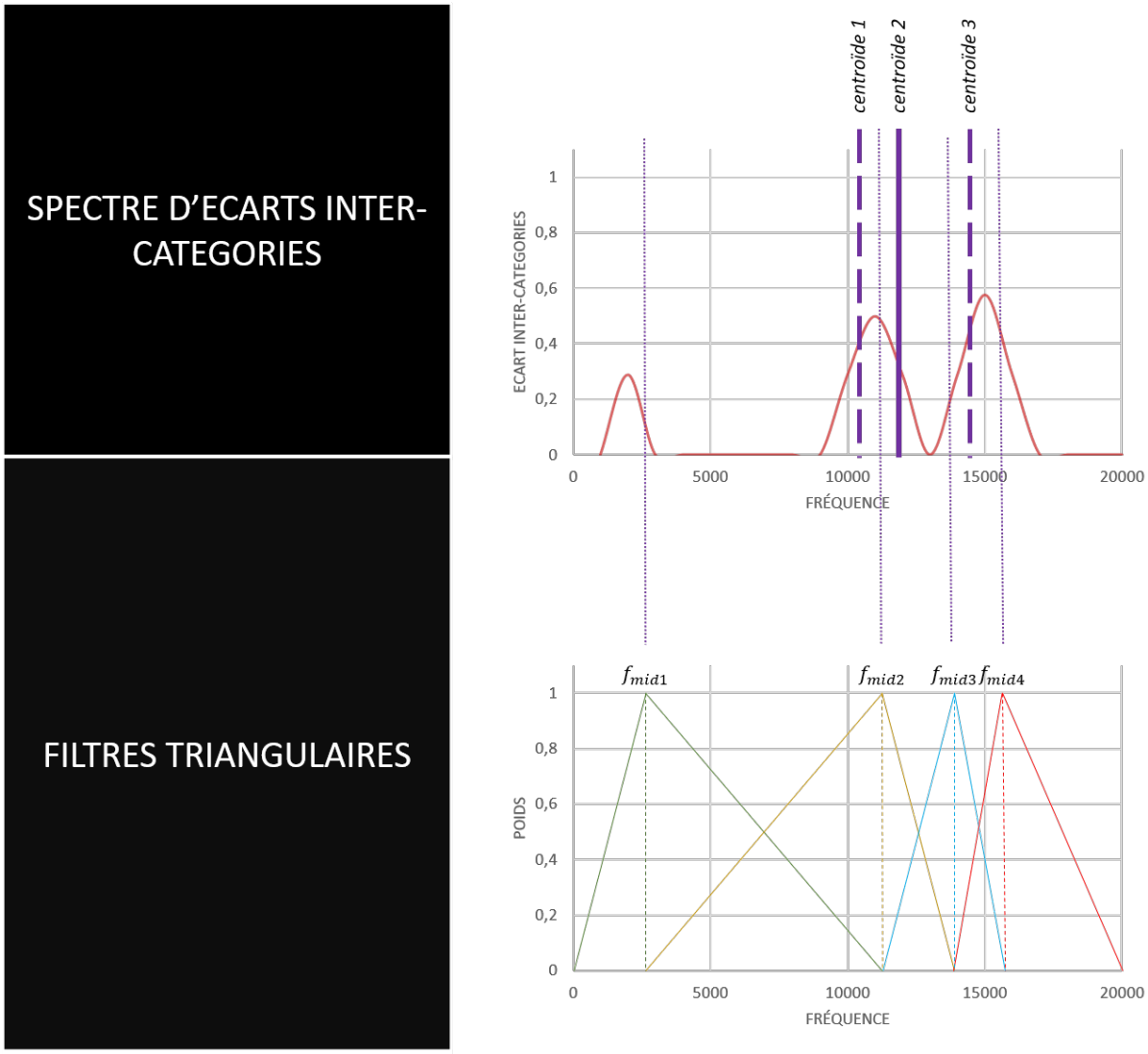

Figure 2.3 – Création d'une banque de 4 filtres à partir du spectre d'écarts inter-catégories

Extraction du contraste spectral

Le contraste spectral est extrait avec un nombre de bandes fixé à 8 et une fréquence minimale de 200 Hz. Une mise en œuvre classique du contraste spectral est d'abord testée, c'est-à-dire en prenant des bandes spectrales correspondant aux octaves de la fréquence 200 Hz.

Dans un second temps, le contraste spectral est calculé pour les bandes de fréquences obtenues à l'issue du calcul présenté dans la méthode DAFCC (voir section 2.3.2).

2.3.3 Modèles de *Machine Learning* entrainés

Entraînement du perceptron multicouche

Le perceptron multicouche est testé avec différentes sous-couches de neurones. Géron (2019) précise que, pour la plupart des problèmes, seulement une ou deux couches intermédiaires sont nécessaires pour obtenir des résultats satisfaisants, avec des couches intermédiaires possédant le même nombre de neurones. Deux couches intermédiaires sont donc choisies pour ce modèle, et chaque couche contient un nombre variable de neurones. Plusieurs modèles sont créés avec un nombre différent de neurones dans les couches intermédiaires pour chaque modèle. Le taux d'apprentissage est fixé à 10^{-3}. L'algorithme d'optimisation ADAM ((Kingma et Ba, 2015)) est utilisé pour la descente de gradient, avec les hyper-paramètres $\beta 1$ et $\beta 2$ fixés à respectivement 0.99 et 0.9. Une régularisation L2 (Hoerl et Kennard, 1970) est utilisée, avec un coefficient de pénalité α fixé à 10^{-2}. La fonction d'activation employée pour chaque neurone est la fonction logistique. Les mini-batch contiennent chacun 200 observations. Le maximum d'itérations de l'algorithme de descente de gradient est fixé à 200. La tolérance est de 10^{-2}, c'est-à-dire que si la fonction coût ne s'améliore pas de plus de 10^{-2} pendant 10 itérations, l'algorithme s'arrête.

Entraînement de la forêt d'arbres décisionnels

Plusieurs modèles de forêt d'arbres décisionnels sont testés, en faisant varier la profondeur maximale d des arbres et le nombre N d'arbres par forêt. Le nombre de descripteurs utilisé dans chaque arbre est égal à la racine carrée du nombre total de descripteurs. Ces descripteurs sont alors sélectionnés aléatoirement parmi l'ensemble des descripteurs. Une technique de *bagging* est utilisée, c'est-à-dire que pour un nombre total d'observations N, N observations sont tirées aléatoirement pour chaque arbre avec remplacement. Chaque arbre contiendra donc environ $\lim_{N\to\infty}(1+1/N)^N \approx 63\%$ des données totales.

Entraînement de la classification naïve bayésienne avec loi gaussienne

La classification naïve bayésienne, lorsque la loi de probabilité gaussienne est supposée, ne possède pas d'hyper paramètre qui pourrait être modifié afin d'améliorer les prédictions. Elle peut cependant être comparée aux autres méthodes réputées pour être plus efficaces, pour s'assurer d'avoir réglé les paramètres des autres méthodes de façon optimale.

2.4 Conclusion pour la partie Méthodes

Une nouvelle taxonomie a été établie à partir des taxonomies existantes et à partir de l'histoire des sous-genres du *heavy metal*. Des enregistrements de chanteurs sont donc effectués pour plusieurs voyelles, pour les techniques que les chanteurs savent pratiquer. A partir de ces enregistrements plusieurs modèles de *Machine Learning* sont développés en extrayant les MFCC, les DAFCC et le contraste spectral. Les DAFCC sont des coefficients inspirés des MFCC, dont les fitres sont adaptés aux données de l'étude.

La section suivante présente la base de données constituée, et les résultats des différents modèles de *Machine Learning* pour cette base de données.

Chapitre 3

Analyse des résultats

3.1 Tri de la base de données

Dans un premier temps, un certain nombre de fichiers audio ont été retirés de la base de données, car ils ont été considérés comme étant trop peu représentatifs de la catégorie attendue. Ce tri aurait pu se faire sur la base des notes issues de l'auto-évaluation des chanteurs (voir section 2.2.4), mais certains chanteurs, bien que ne pratiquant que très rarement une technique, pouvaient la maîtriser assez pour que leurs enregistrements puissent être pris en compte. Cette méthode permet donc de ne garder que les données les plus cohérentes. Finalement, environ 24% des données ont été retirées de la base de données de cette façon.

Suite à cette réduction, la base de données est constituée de 725 fichiers audio. Après découpage de chaque fichier en fenêtres de 1024 échantillons avec un recouvrement de moitié, le nombre d'observations s'élève à 511623. Le nombre de *tunnel throat (TT)*, *deep gutturals (DeG)* et *pig squeal (PS)* est très faible (respectivement 0,3%, 0,9% et 0,7% du total des données) et ces effets n'ont été enregistrés que par respectivement 3 sujets, 8 sujets et 7 sujets (voir figure 3.1 et 3.2). La technique de *grind inhale (GI)* est également très peu représentée (4,1 % des données, et seulement 7 chanteurs). Par conséquent, les enregistrements issus de ces techniques vocales ne seront pas exploités par la suite : ne seront gardées dans cette étude que les techniques de voix claire (CV), de *hardcore scream* (HS), de *black shriek* (BS) et de *death growl* (DG). Les autres techniques pourront être étudiées dans la phase de test, afin d'avoir un aperçu de la catégorie dans laquelle l'algorithme de *Machine Learning* choisi pourrait classer ces données.

Comme le présente la figure 3.3, le nombre d'observations au sein des catégories restantes est assez déséquilibré, la technique de *black shriek* correspondant à notamment environ trois fois moins d'observations que la catégorie voix claire. Plusieurs algorithmes, avec et sans méthodes d'*undersampling* seront testés, afin d'éviter le biais qui peut être engendré par un déséquilibre des données (voir partie 1.5.1).

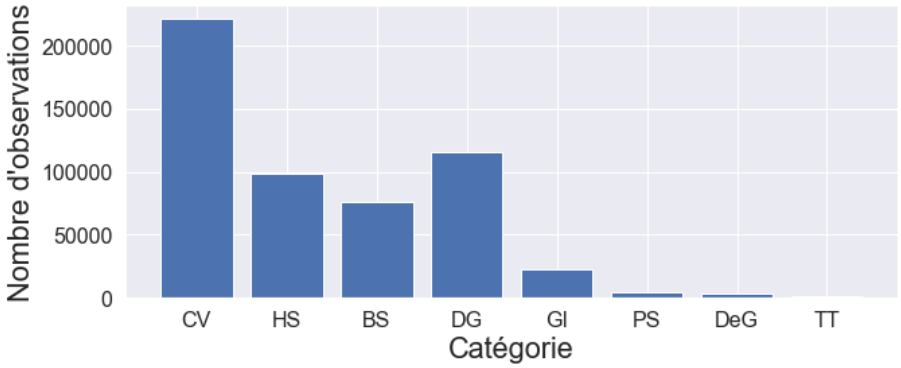

Figure 3.1 – Nombre d'observations par effet ou catégorie.

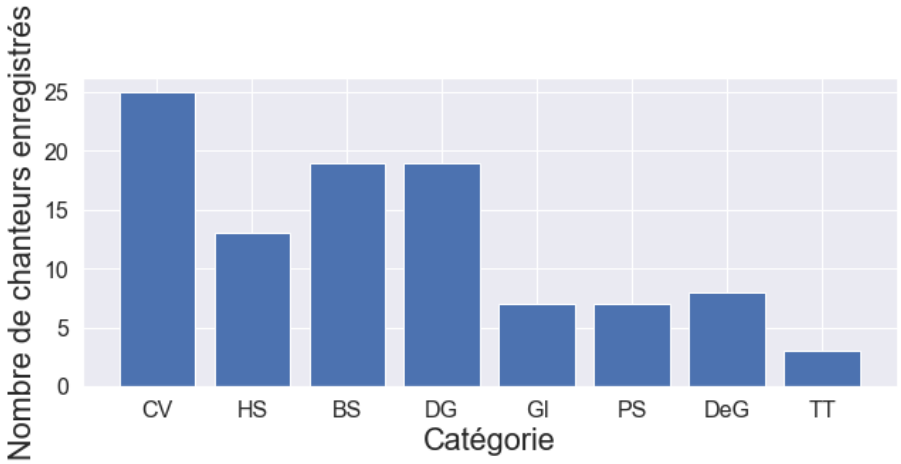

Figure 3.2 – Nombre de chanteurs ayant été enregistrés pour chaque effet ou catégorie.

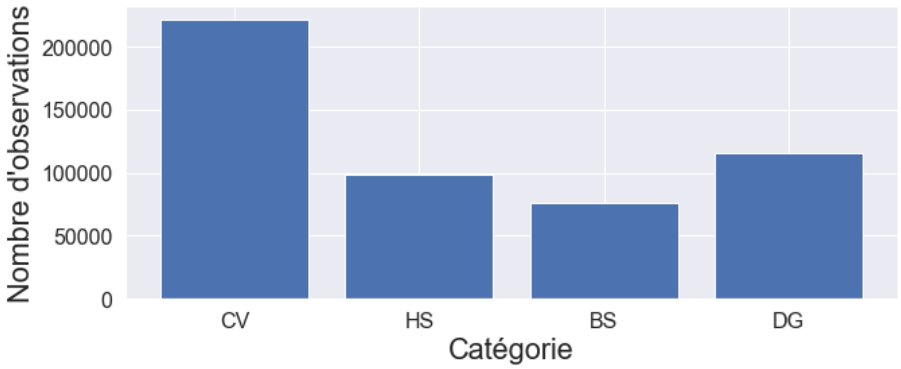

Figure 3.3 – Nombre d'observations par catégories gardées pour l'étude.

3.2 Analyse acoustique des différentes catégories

Les enregistrements, pour les voyelles [a], [i] et [u] montrent très nettement des différences entre le chant clair et le chant saturé (voir figure 3.4, 3.5 et 3.6) du fait de la présence très marquée de bruit dans le spectrogramme de chacune des techniques vocales extrêmes. Il est ainsi très probable que le score de précision d'un modèle de *Machine Learning* pour la technique de chant clair soit meilleur que ceux des autres techniques. Des phénomènes de *shimmer* et de *jitter* sont observables sur toutes les techniques de chant saturé. Les techniques de *black shriek* et de *death growl* sont bruitées sur une grande partie du spectre (entre 50 Hz et 5 kHz), tandis que le *hardcore scream* est bruité dans une zone plus petite et centrée sur la fréquence fondamentale. En ne se fiant qu'à l'enveloppe du signal et qu'au spectrogramme, il est très compliqué de différencier le *black shriek* du death growl, car ils possèdent tous deux des propriétés acoustiques très proches. Cette particularité devrait également se ressentir dans la matrice de confusion, car il est probable que beaucoup de *black shriek* soient confondus avec les *death growl* et inversement.

Dans l'enveloppe du *death growl* présentée en figure 3.4, on peut observer une macro impulsion (en rouge) qui regroupe plusieurs impulsions glottales (en vert) caractéristiques de techniques de chant saturé, comme avait pu le constater Nieto (2008). Les deux premières impulsions glottales de la macro impulsion présentée sont très marquées, la troisième dispose d'un niveau beaucoup plus faible. Dans cet exemple, la fréquence de macro impulsion semble apparaître sur le spectrogramme avec une bande très légèrement dessinée aux alentours de 70 Hz (donc pour une période de 15 ms sur la figure 3.4). L'impulsion glottale, plus marquée, est proche des 200 Hz (donc pour une période de 5 ms sur la figure 3.4). L'impulsion glottale de la voix claire, d'une période d'environ 5 ms, a également été représentée en vert dans la figure 3.4 à des fins de comparaison. Dans la plupart des enregistrements, la saturation engendrée est trop irrégulière, c'est-à-dire que le *jitter* et le *shimmer* sont trop élevés (voir partie 1.4), pour percevoir un effet significatif de macro impulsions.

Aucune différence acoustique significative ne peut être observée dans l'utilisation des voyelles entre le *death growl* et le *black shriek* : les spectrogrammes du *black shriek* et du *death growl* sont quasiment identiques (voir figures 3.8 et 3.9). En revanche, la voyelle [u], pour chacune des trois techniques saturées, est bruitée dans une zone spectrale beaucoup plus basse que les autres voyelles (avec une énergie spectrale globalement regroupée en dessous de 1 kHz). La technique de *hardcore scream* prononcée avec un [u] (voir figure 3.7), possède notamment un spectrogramme très proche de celui du *death growl* et du black shriek. Ainsi, il est possible qu'un algorithme de *Machine Learning* confonde les voyelles [u] du *hardcore scream* avec des techniques de *death growl* ou de *black shriek*. Plus largement, un *hardcore scream* dans le registre *low* pourrait être plus facilement confondu avec le *black shriek* ou le *death growl*.

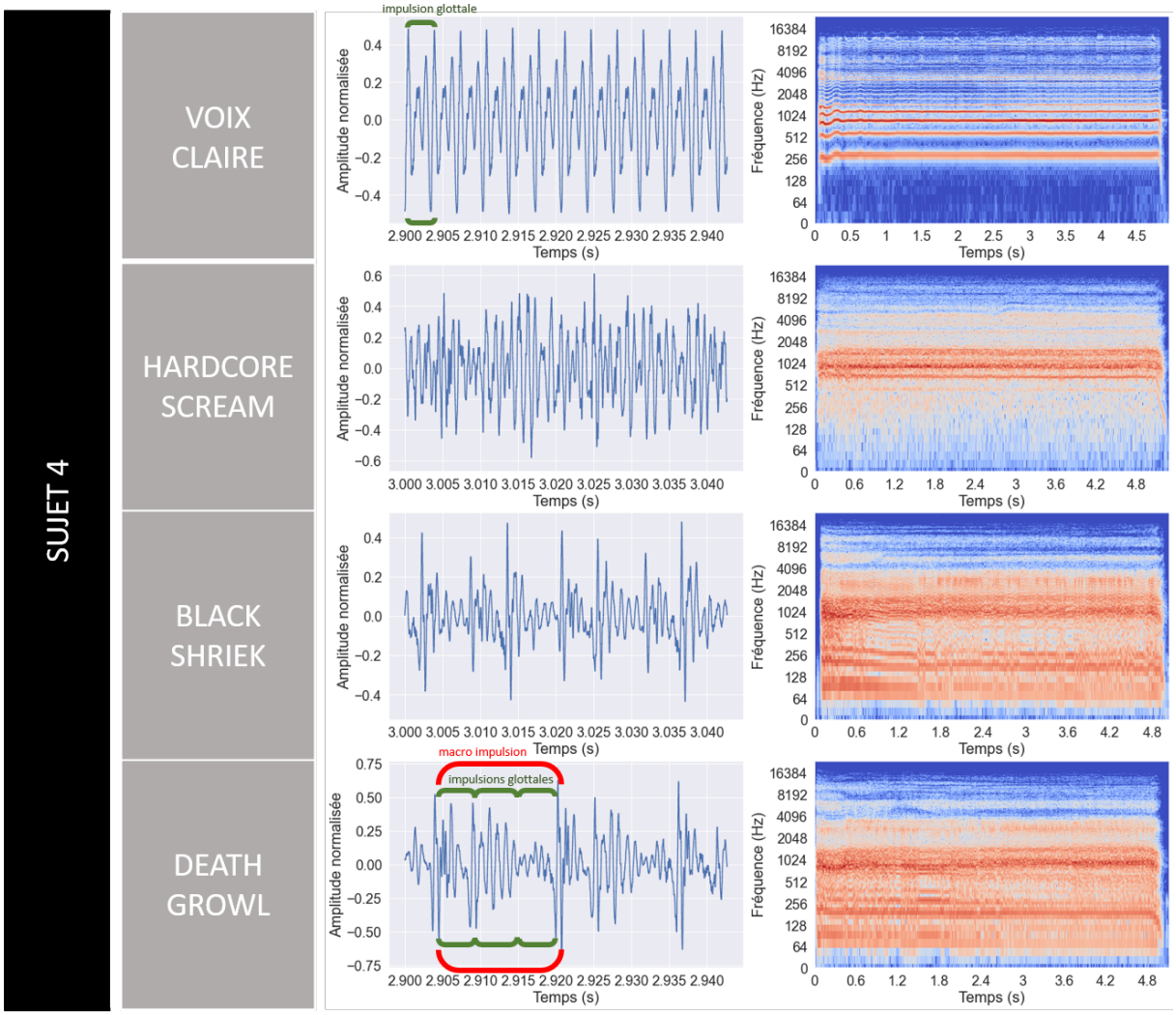

Figure 3.4 – Enveloppe du signal et spectrogramme pour les différentes techniques employées par le sujet 4 lors de la production de la voyelle [a] dans le registre *mid*.

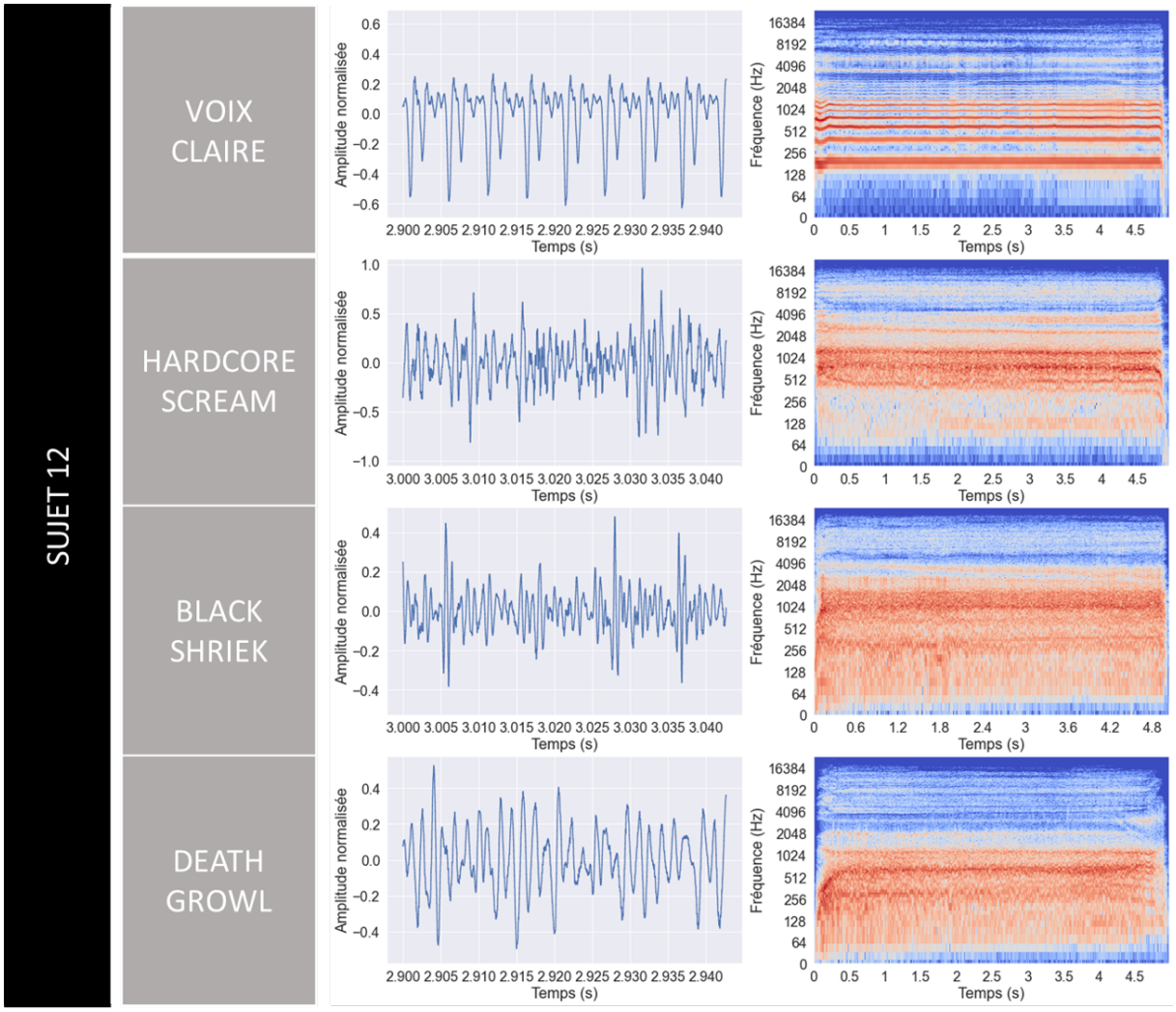

Figure 3.5 – Enveloppe du signal et spectrogramme pour les différentes techniques employées par le sujet 12 lors de la production de la voyelle [a] dans le registre *mid*.

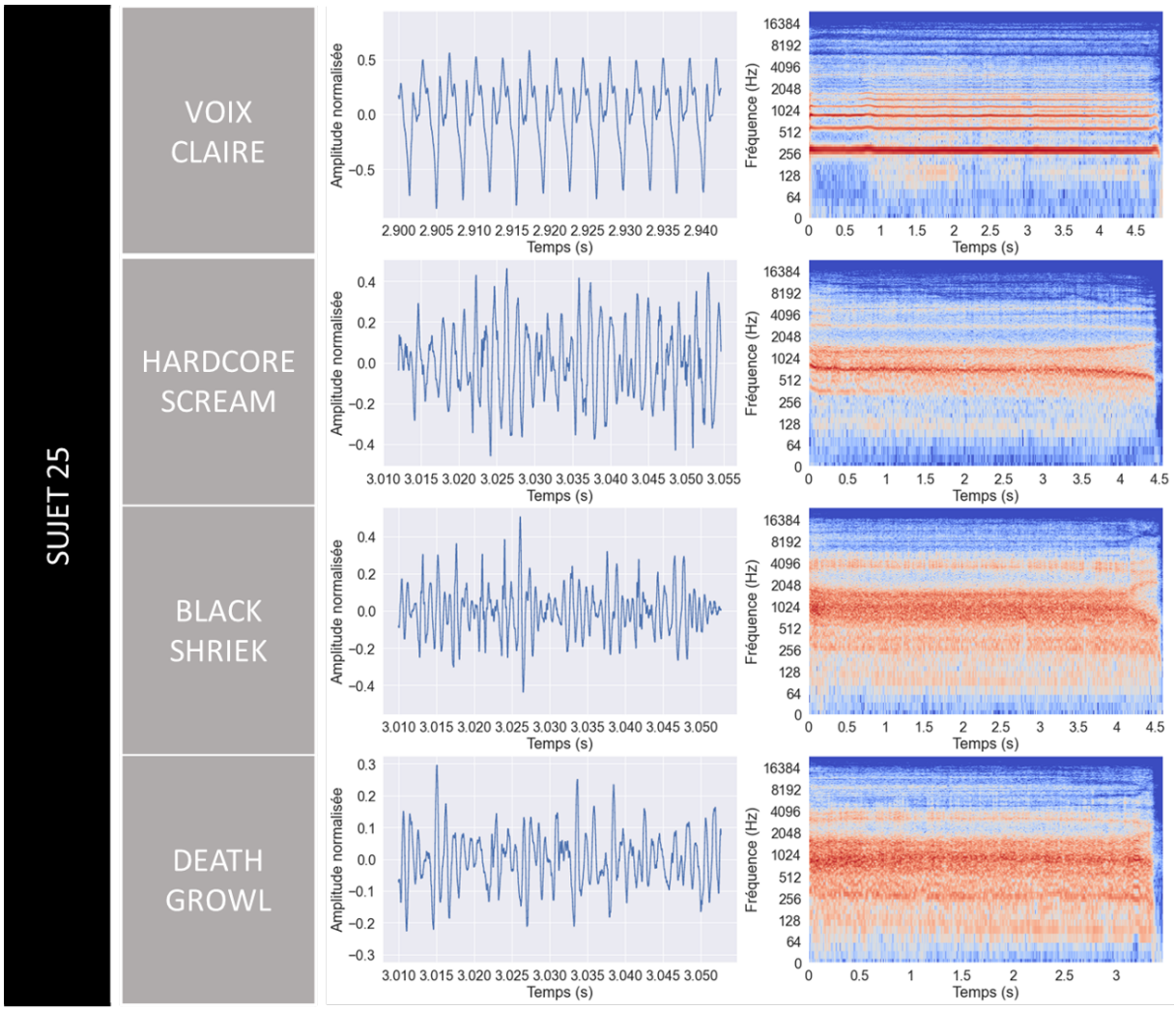

Figure 3.6 – Enveloppe du signal et spectrogramme pour les différentes techniques employées par le sujet 25 lors de la production de la voyelle [a] dans le registre *mid*.

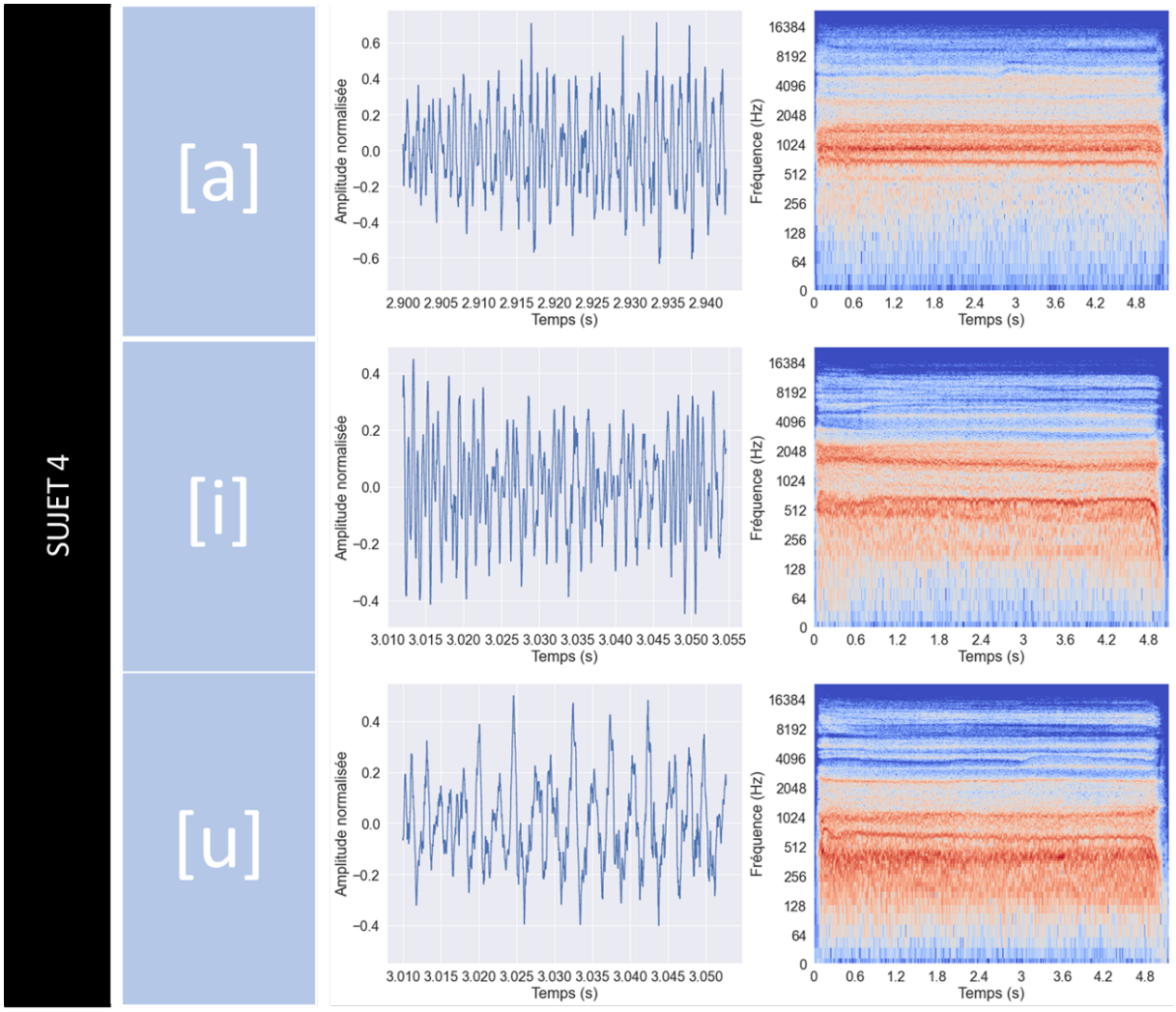

Figure 3.7 – Enveloppe du signal et spectrogramme pour les différentes voyelles employées par le sujet 4 lors de la production d'un *hardcore scream* dans le registre *mid*.

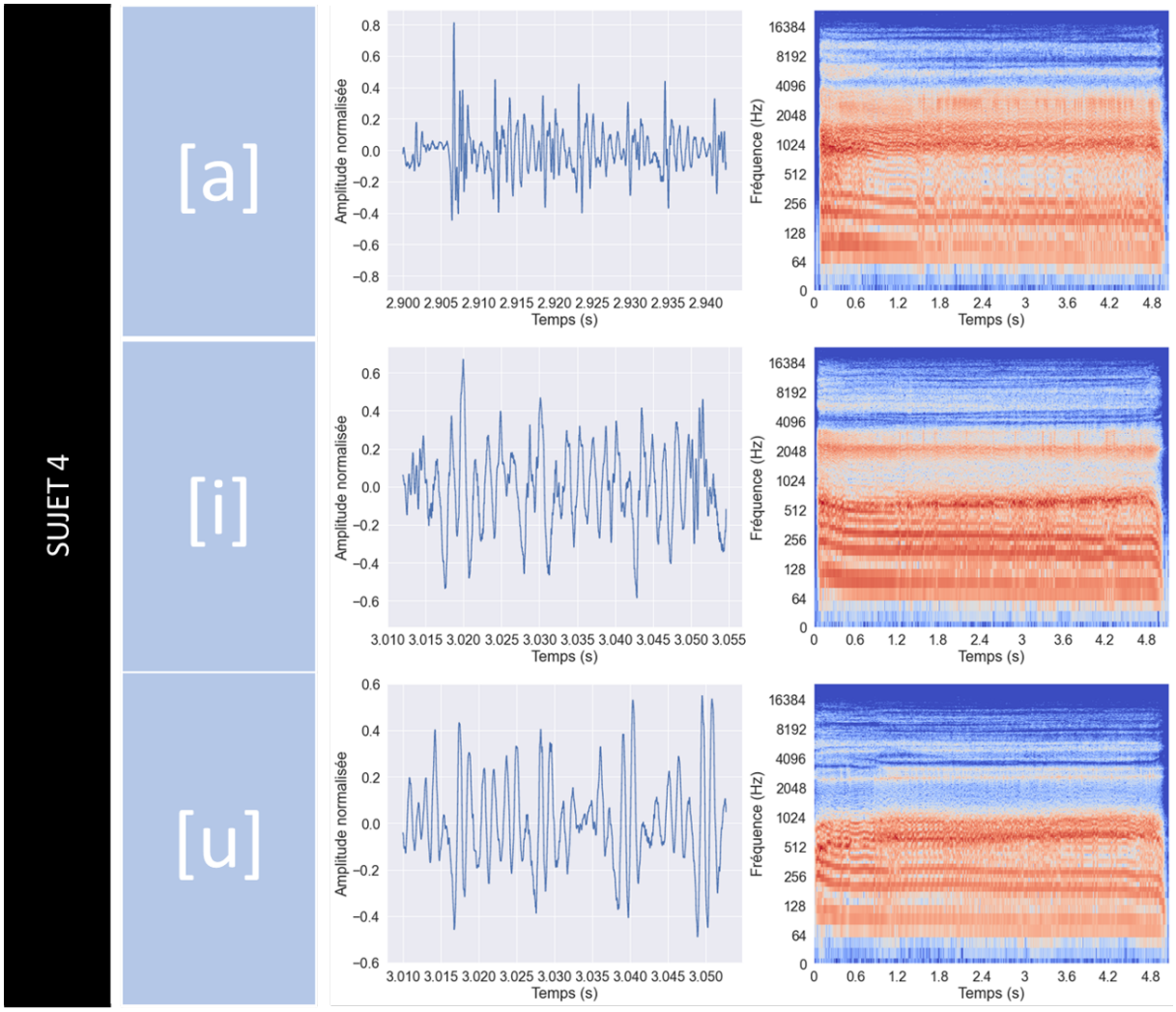

Figure 3.8 – Enveloppe du signal et spectrogramme pour les différentes voyelles employées par le sujet 4 lors de la production d'un *black shriek* dans le registre *mid*.

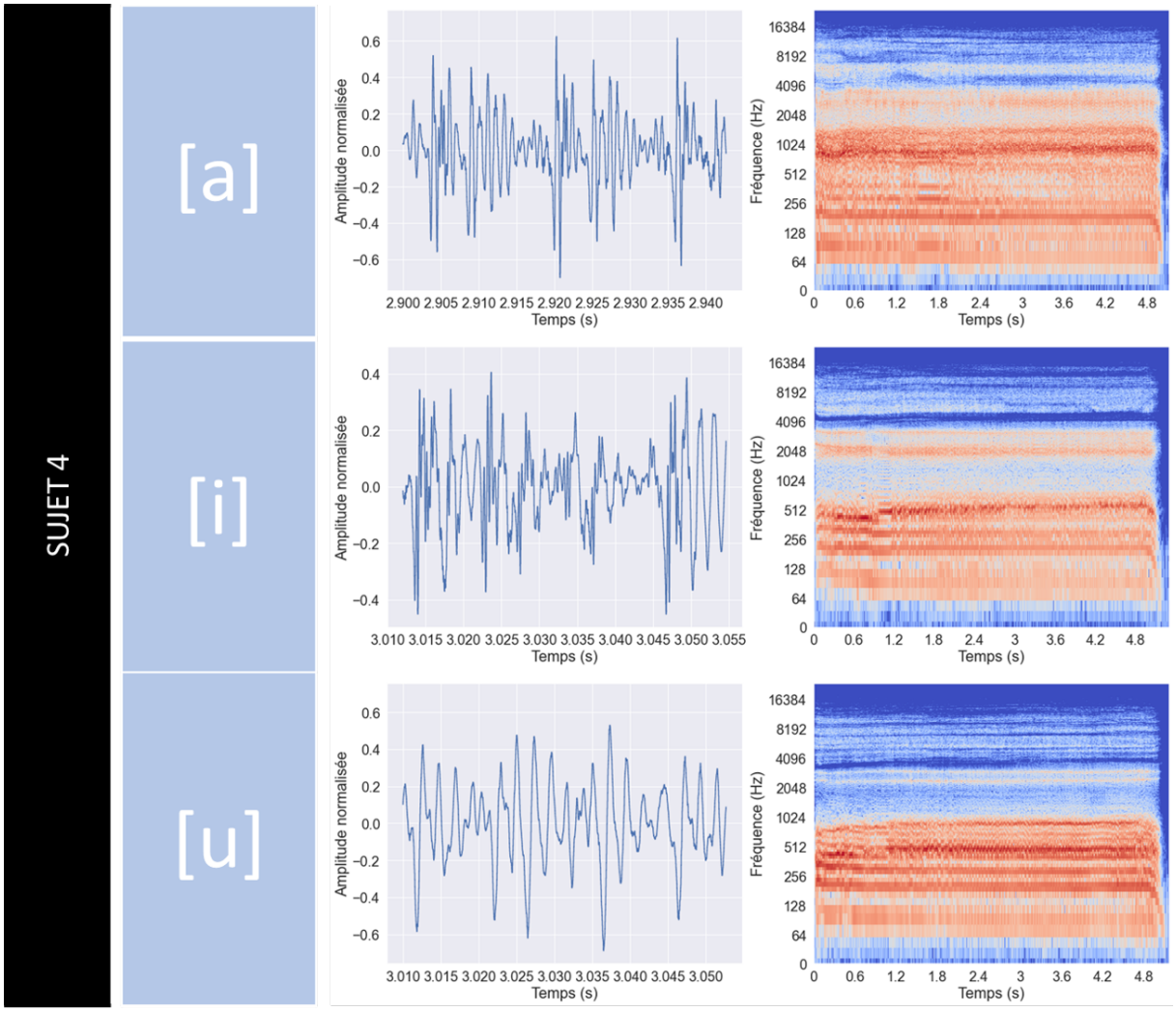

Figure 3.9 – Enveloppe du signal et spectrogramme pour les différentes voyelles employées par le sujet 4 lors de la production d'un *death growl* dans le registre *mid*.

3.3 Analyse de plusieurs algorithmes : choix d'un modèle et de descripteurs

Dans les sections 3.3.1, 3.3.2 et 3.3.3, les descripteurs utilisés sont les 13 premiers MFCC avec 128 filtres de Mel générés à partir de la formule de Slaney. Ces descripteurs ne seront modifiés qu'en section 3.3.4.

3.3.1 Influence d'un *undersampling* aléatoire sur la précision

Afin de valider ou non l'utilisation de méthodes d'*undersampling*, trois modèles ont été testés avec l'utilisation ou non d'un *undersampling* aléatoire (voir figure 3.10).

Avec cet *undersampling*, le nombre d'observations utilisées pour l'entraînement du modèle est ramené à 76179 pour chaque classe. Ce nombre correspond au nombre d'observations dans la catégorie *black shriek*, catégorie minoritaire. Rappelons ici que l'*undersampling* n'est utilisé que sur les données d'entraînement, mais que l'algorithme est testé sur toutes les observations. Les matrices de confusion de chaque modèle montrent qu'en l'absence d'undersampling, les différents modèles deviennent sur-entraînés dans la détection de voix claire, et moins performants pour les techniques de chant saturé. Cet effet est particulièrement visible pour la forêt d'arbres décisionnels, avec une précision de 91% pour la voix claire, mais de seulement 36% et 28% pour le *hardcore scream* et le *black shriek*. De plus 45% des *hardcore scream* et 43% des *black shriek* ont été identifiés comme des voix claires, ce qui laisse penser que le problème vient effectivement d'une sur-représentation de la catégorie voix claire dans la base de données d'entraînement. Un *undersampling* aléatoire sera donc utilisé lors de la phase d'entraînement.

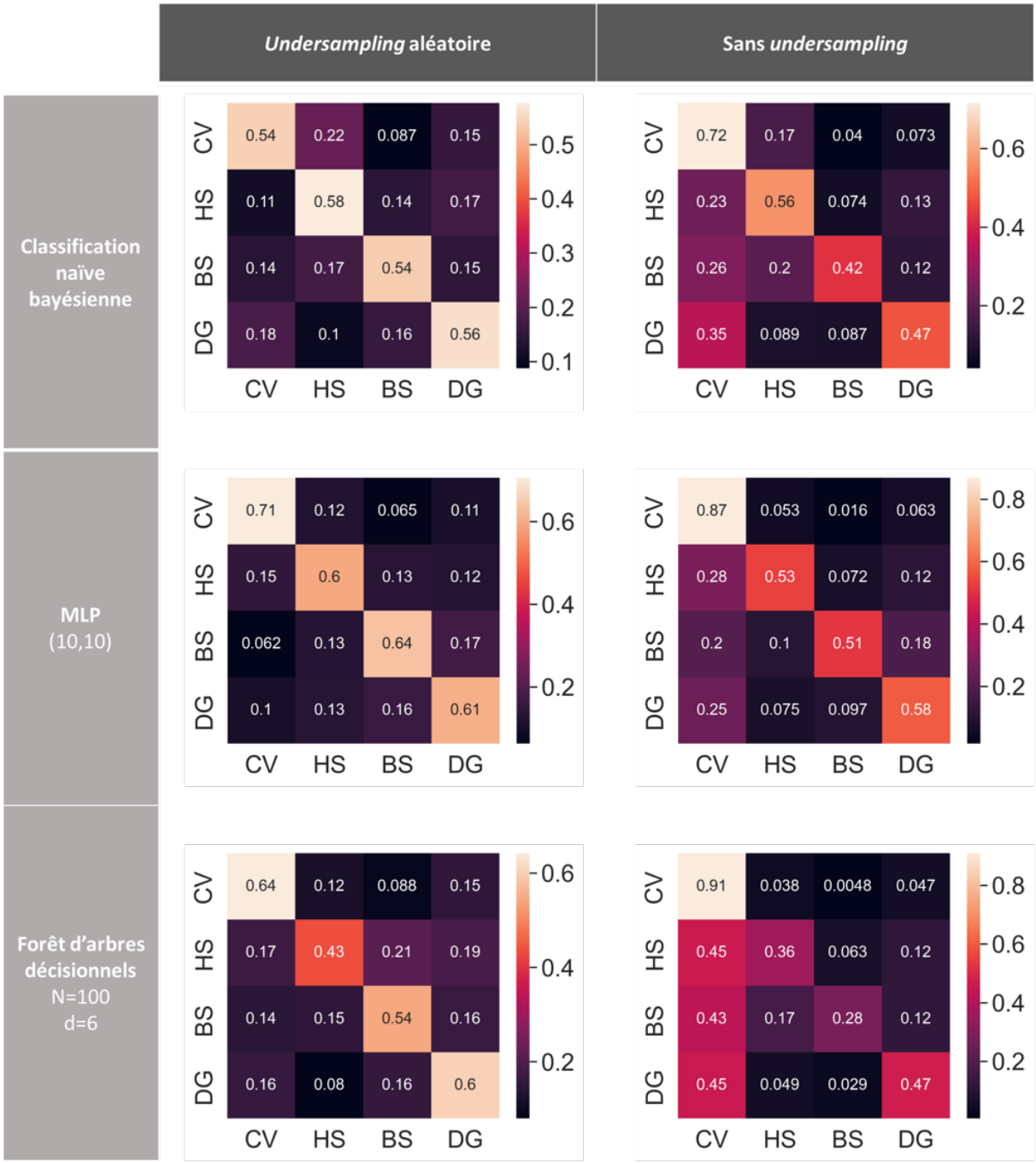

Figure 3.10 – Matrice de confusion avec et sans *undersampling* pour plusieurs modèles.

3.3.2 Analyse comparative des modèles de *Machine Learning*

Dans un premier temps, l'algorithme de forêt d'arbres décisionnels a été testé en fixant la profondeur maximale à 6 et en augmentant le nombre d'arbres à 200 ou 300. Pour ces différents modèles, la précision stagne entre 55,60% et 55,65% (voir figure 3.11). Le même phénomène se produit en faisant varier le nombre d'arbres de la forêt pour une profondeur maximale de 10, la précision restant aux alentours de 61,1%. En fixant le nombre d'arbres de la forêt à une valeur de 100, l'augmentation de la profondeur maximale permet d'augmenter beaucoup plus significativement la précision (voir figure 3.12). La précision passe ainsi de 55,6% pour d=6, à 63,9% pour d=15. Les paramètres N=100 et d=15 semblent ainsi donner les meilleurs résultats pour la forêt d'arbres décisionnels.

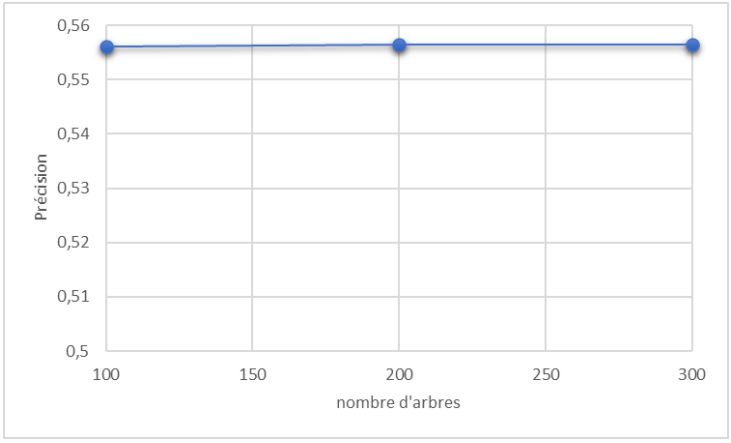

Figure 3.11 – Précision en fonction du nombre d'arbres dans la forêt d'arbres décisionnels, avec une profondeur de 6.

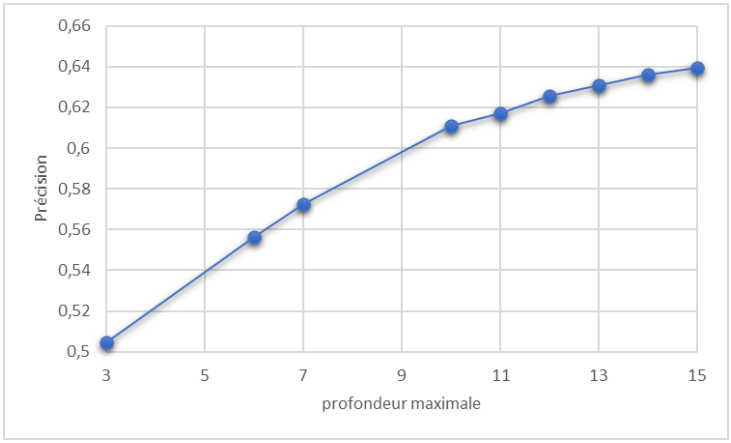

Figure 3.12 – Précision en fonction de la profondeur maximale des arbres de la forêt décisionnelle, avec un nombre d'arbres de 100.

L'algorithme de classification naïve bayésienne obtient une précision de 55,4%.

Pour le perceptron multicouche, le maximum de précision de 64,7% est atteint lorsque l'algorithme exploite 200 neurones pour chacune des deux couches intermédiaires (voir figure 3.13).

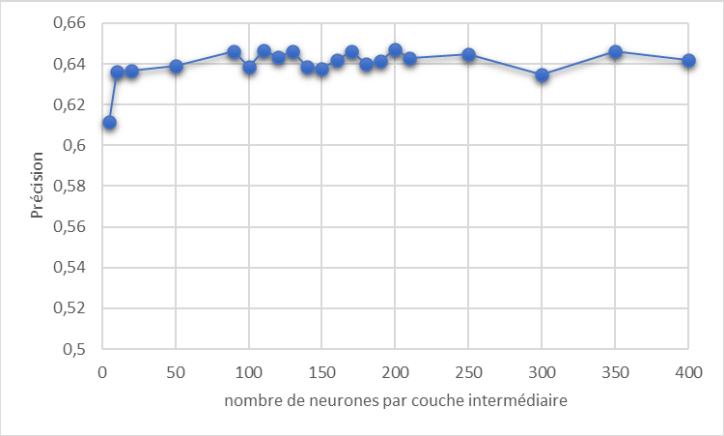

Figure 3.13 – Précision en fonction du nombre de neurones par couches intermédiaires.

La figure 3.14 montre les matrices de confusions des trois modèles, les paramètres de chaque modèle étant réglés de façon à obtenir une précision maximale. La classification naïve bayésienne est très clairement la moins efficace pour la classification, chaque classe obtenant une précision moins bonne que pour les deux autres modèles. Bien que la forêt d'arbres décisionnels obtienne des résultats très légèrement supérieurs à ceux du perceptron multicouche pour les catégories de *death growl* et de *hardcore scream*, les catégories *hardcore scream* et voix claire obtiennent un moins bon résultat. Le perceptron multicouche obtient au total une précision de 64,7% contre 63,9% pour la forêt d'arbres décisionnels. Ces résultats sont très proches, mais le perceptron multicouche sera utilisé pour la suite, car à performances équivalentes, le temps de calcul du perceptron multicouche lors de la phase d'entraînement est environ trois fois plus court que celui de la forêt d'arbres décisionnels.

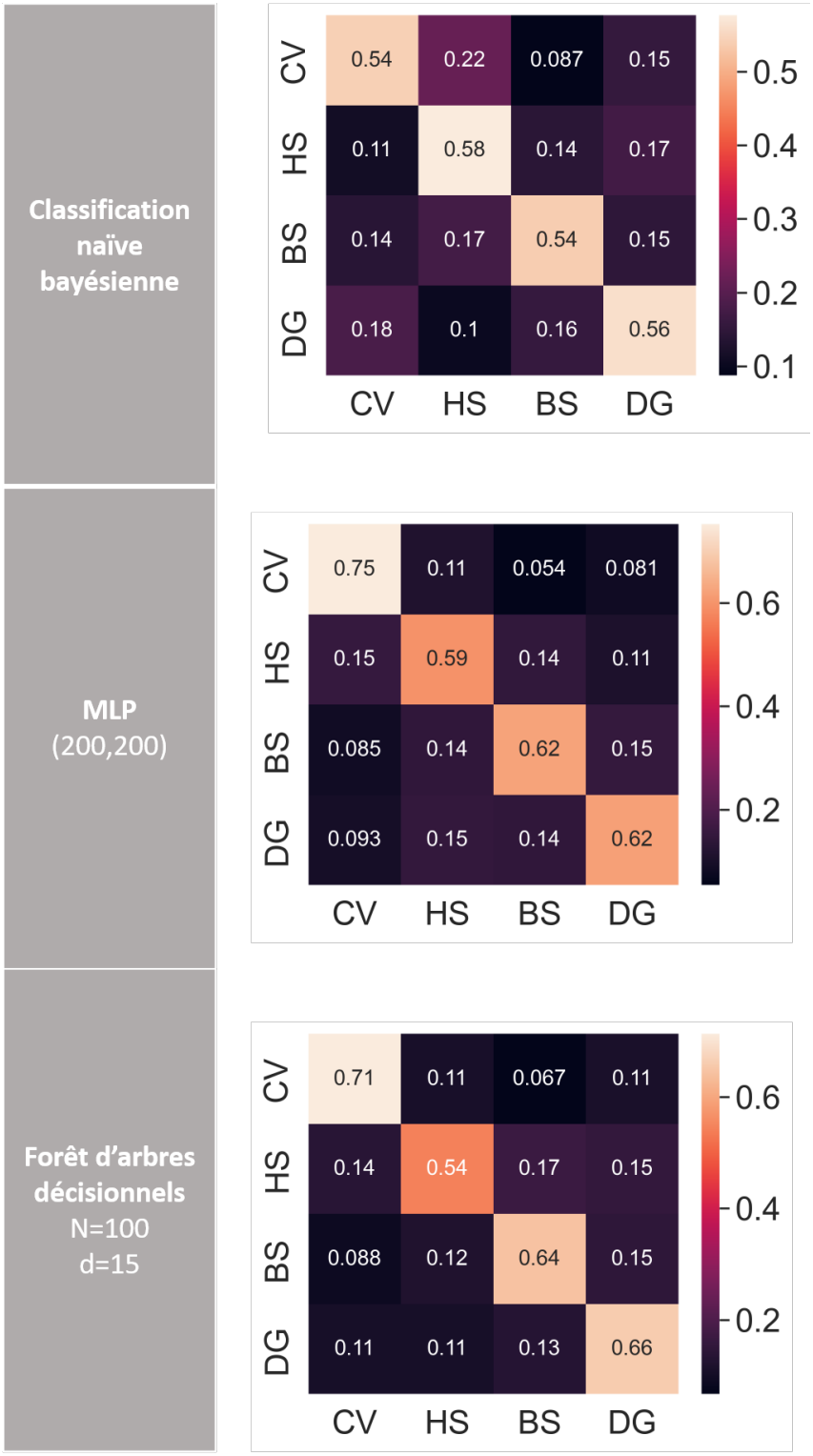

Figure 3.14 – Matrices de confusion des meilleurs modèles de classification naïve bayésienne, de perceptron multicouche, et de forêt d'arbres décisionnels.

3.3.3 Suppression des silences

Les extraits sonores étudiés, particulièrement ceux basés sur l'enregistrement de paroles, peuvent contenir des respirations et des silences. Ces silences sont donc à la fois pris en compte pour les données d'entraînement, mais également pour les données de test, alors qu'ils ne représentent aucune catégorie. Pour pallier cela, la RMSE (Root Mean Square Energy) (voir équation 3.1) est calculée pour chaque trame étudiée. $s = (s_1, s_2, ..., s_N)$ désigne ici le signal normalisé, contenant N échantillons (donc 1024 pour cette étude). Si la RMSE dépasse un certain seuil t_k, alors les données sont utilisées lors des phases d'entraînement et de test. Dans le cas contraire, ces données ne sont tout simplement pas étudiées par l'algorithme.

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N} s_k^2} \tag{3.1}$$

Cette méthode reste cohérente dans l'optique du développement d'un plug-in qui pourrait être utilisé sur scène : les données ne dépassant pas le seuil pourraient, par exemple, être étiquetées dans la classe la plus utilisée par le chanteur (par exemple la classe voix claire).

En supprimant les données dont la RMSE est inférieure à 10^{-3}, seuls les silences entre les phrases ne sont pas pris en compte pour les analyses, cela ne modifie donc que très peu les données (voir figure 3.15). Les résultats de l'algorithme deviennent alors significativement meilleurs, la précision du perceptron multicouche passant de 64,7% à 68,5%.

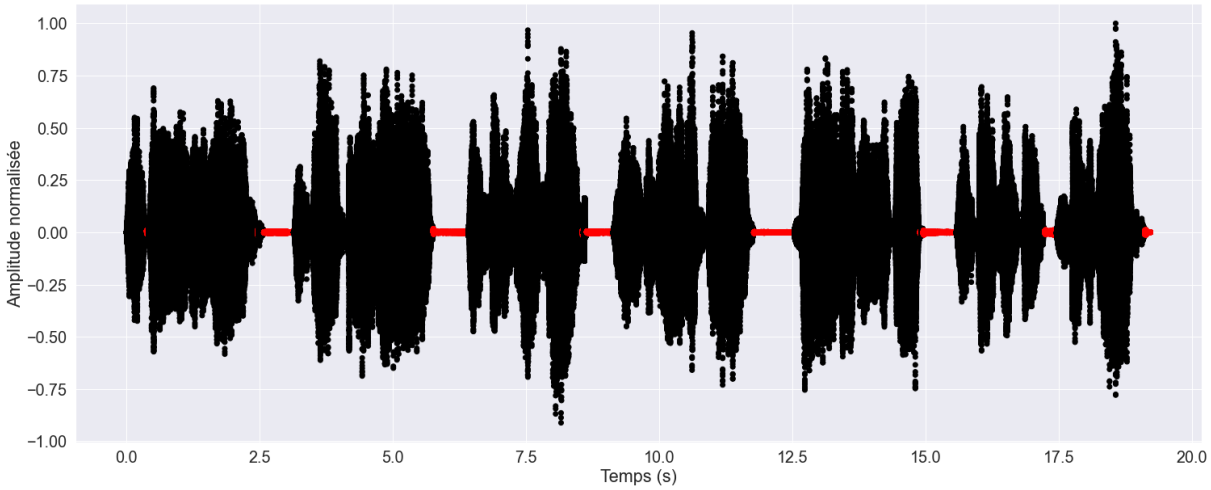

Figure 3.15 – Exemple de calcul de la RMSE pour un fichier audio. Les données en noir dépassent le seuil de 10^{-3}, celles en rouge sont en dessous de ce seuil.

3.3.4 Analyse comparative des descripteurs

Dans cette section, le modèle sera fixe (perceptron multicouche choisi en section 3.3.2), et plusieurs descripteurs seront testés.

Beaucoup d'études conseillent de retirer le premier MFCC des descripteurs, qui est corrélé à la puissance moyenne du spectre et peut donc être remplacé par un calcul de puissance (Zelkowitz, 2010). En revanche, pour cette étude, retirer ce coefficient, toujours dans le contexte d'une extraction de 13 MFCC et avec des filtres conçus à partir de la formule de Slaney, conduit à une baisse conséquente de la précision, qui passe de $68,5\%$ à $66,9\%$. Cela peut s'expliquer par le fait que les techniques de chant saturées possèdent une puissance spectrale beaucoup plus importante que la voix claire (voir par exemple la figure 3.6). En remplaçant le premier coefficient par le calcul de la puissance spectrale (voir équation 1.4), la précision baisse de 68,5% à 67,1%. Le premier coefficient des MFCC sera donc conservé. Afin d'améliorer la précision, il est possible de changer le nombre de MFCC à extraire, ou de changer la formule faisant la conversion entre les hertz et les mels (Slaney ou Young). Un maximum de 128 MFCC peuvent être utilisés comme descripteurs (puisque le nombre de filtres de Mel a été fixé à 128). La figure 3.16 montre l'évolution de la précision du modèle en fonction du nombre de MFCC conservés pour l'utilisation des formules de Slaney et de Young. Le maximum de précision de 74,5% est atteint pour un nombre de MFCC de 33 en utilisant la formule de Young. En gardant les 33 premiers MFCC, baisser le nombre de filtres de Mel conduit à une baisse de précision (voir figure 3.17). En utilisant 24 filtres de Mel et en gardant les 24 premiers MFCC, la précision descend jusqu'à 72,1%. Le nombre de filtres de Mel sera donc maintenu à 128. La soustraction cepstrale ne permet pas d'obtenir de meilleurs résultats, faisant même baisser la précision de 74,5% à 74,2%. Il est important de préciser ici que la prédiction peut beaucoup fluctuer en ajoutant ou retirant un seul MFCC, et qu'un extremum de précision plus élevé existe peut-être pour une autre valeur non représentée dans ces différentes figures.

L'extraction des 8 coefficients du contraste spectral standard, obtenus à partir des 7 octaves de 200 Hz, conduit à un score de précision de seulement 54,6%. L'extraction des 16 coefficients de contraste spectral obtenus à partir des 16 bandes spectrales utilisées pour le calcul des DAFCC conduit à une précision de 58,9%. Ces résultats, malgré le faible nombre de coefficients utilisés, sont bien en dessous du maximum de précision de 74,5% obtenus à l'issue de l'extraction des MFCC.

Avec 18 coefficients DAFCC (et 32 filtres), le modèle atteint une précision de 75,2%, ce qui est même meilleur que le score qu'avaient obtenu les MFCC pour 33 coefficients (et 128 filtres). Ce modèle permet ainsi de réduire à la fois le nombre de descripteurs, mais également le nombre de filtres, tout en améliorant la précision. La figure 3.18 montre l'évolution de la précision de l'algorithme en fonction du nombre de DAFCC. Ce graphique montre que la classification basée sur l'utilisation des DAFCC est moins sensible au nombre de coefficients conservés que la classification basée sur l'utilisation des MFCC puisque la précision reste située entre 73,9% et 74,9% pour les combinaisons testées alors qu'elle varie entre 70,2% et 74,5% pour la classification basée sur les MFCC calculés via la formule de Young.

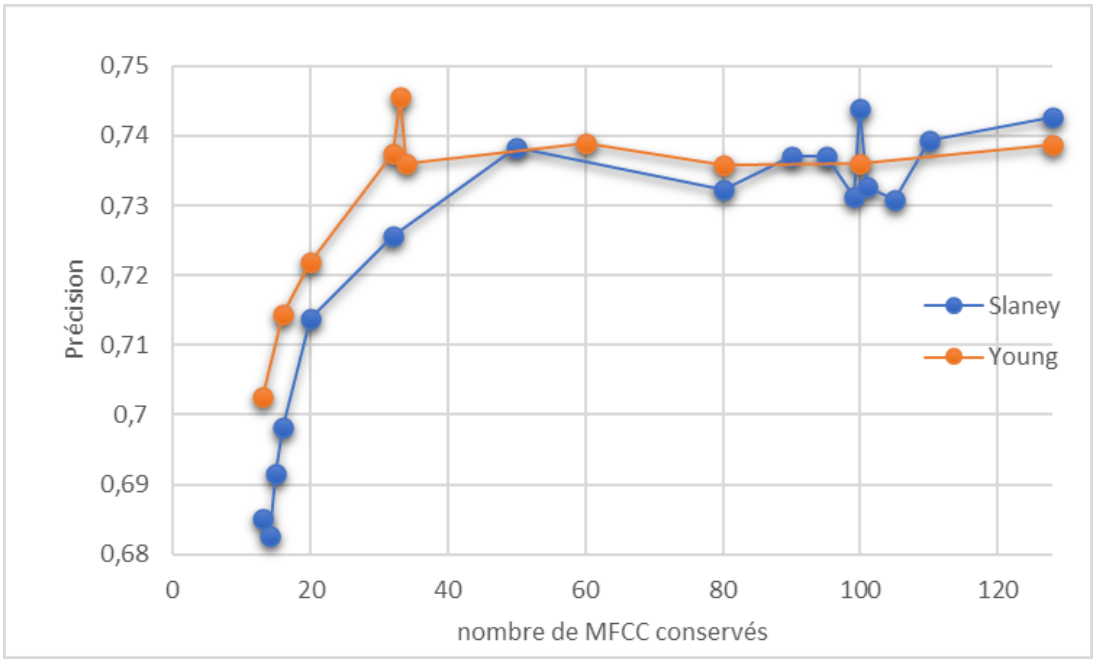

Figure 3.16 – Précision en fonction du nombre de MFCC conservé, avec l'utilisation de la formule de Slayney ou de la formule de Young pour la conversion de hertz à mels.

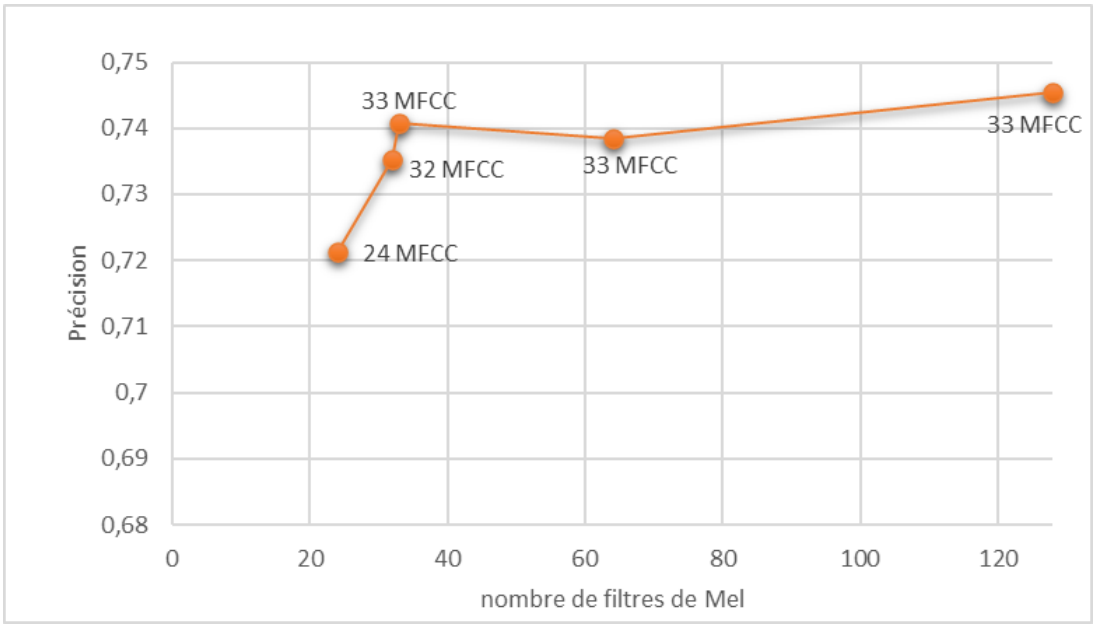

Figure 3.17 – Précision en fonction du nombre de filtres de Mel utilisés, avec l'utilisation de la formule de Young. Le nombre de MFCC conservés est précisé sur chaque point du graphique.

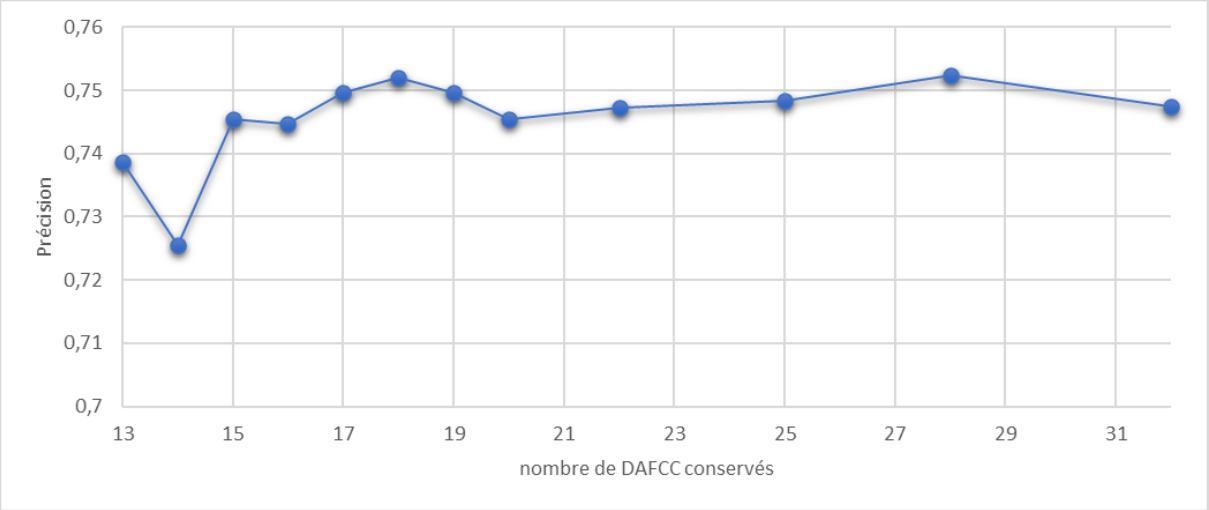

Figure 3.18 – Précision en fonction du nombre de DAFCC conservés pour 32 filtres.

3.3.5 Modification de la taille des trames

Avec 18 coefficients DAFCC (et 32 filtres), le modèle passe d'une précision de 75,2% pour des trames de 1024 échantillons, à 75,3% pour des trames de 2048 échantillons, puis à 74,0% pour des trames de 4096 échantillons. Autrement dit, augmenter la taille des échantillons résulte en une précision du modèle très proche de la précision initiale obtenue pour une trame de 1024 échantillons, voire même une précision inférieure.

La réduction de la taille des trames, qui implique une baisse de la résolution fréquentielle, ne peut être effectuée sans changer le nombre de filtres. En effet, en gardant 32 filtres pour l'étude, certains filtres se retrouvent sans aucune données fréquentielles à analyser. Pour les trames de 512 échantillons, 16 filtres ont donc été utilisés, et 16 DAFCC ont été extraits. La précision est alors de 72,1%. Pour des trames de 256 échantillons, analysés avec 8 filtres, la précision baisse encore et atteint 59,6% (voir figure 3.19).

DAFCC conservés	Nombre de filtres	Trame	Précision
18	32	4096	0,740
18	32	2048	0,753
18	32	1024	0,752
16	16	512	0,721
8	8	256	0,596

Figure 3.19 – Précision en fonction du nombre d'échantillons par trame.

3.4 Analyse du modèle final

Suite aux résultats obtenus dans la section 3.3, les descripteurs extraits sont les 18 premiers DAFCC (avec 32 filtres), pour des trames de 1024 échantillons. Le modèle sélectionné est le perceptron multicouche avec 200 neurones pour chacune des deux couches intermédiaires, avec un *undersampling*, et une suppression des silences.

3.4.1 Résultats globaux du modèle

Ce modèle atteint une précision de 75,2%. La matrice de confusion finale, présentée en figure 3.20, conforte bien les hypothèses faites dans la section 3.2. En effet, la catégorie voix claire se démarque très clairement des autres catégories, avec un taux de détection de 90%. De plus, très peu de *hardcore scream* (3,6%), de *black shriek* (3,1%) et de *death growl* (5,4%) ont été confondus avec la catégorie voix claire. En revanche, les trois catégories peuvent se confondre les unes avec les autres de façon assez significative, avec une confusion variant entre 11% et 16%.

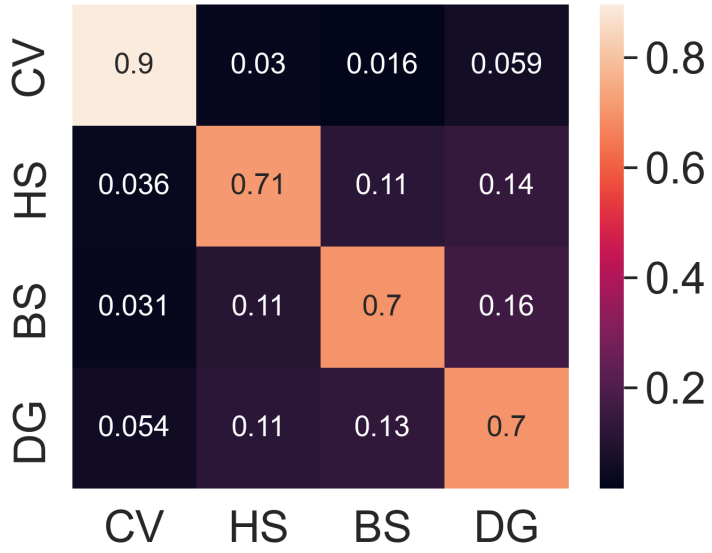

Figure 3.20 – Matrice de confusion du perceptron multicouche avec 200 neurones pour chacune des deux couches intermédiaires, exploitant 18 DAFCC.

3.4.2 Résultats par catégorie et registre

La matrice de confusion détaillée en figure 3.21 montre qu'avec ce modèle, 32% des *hardcore scream* dans le registre *low* ont été étiquetés dans la catégorie *death growl*. De même, 22% des *black shriek* dans le registre *mid* ont été étiquetés à tort dans la catégorie *death growl*. Enfin, 21% des *death growl* dans le registre *mid* ont été étiquetés dans

la catégorie *black shriek*. Ces résultats confirment donc les hypothèses faites dans la section 3.2 : les *black shriek* et les *death growl* se confondent beaucoup dans le registre *mid*, et le *hardcore scream* dans le registre *low* peut facilement être confondu avec un *death growl*.

catégorie	registre	CV	HS	BS	DG
voix claire	High	0,89	0,07	0,01	0,03
	Mid	0,92	0,01	0,02	0,05
	Low	0,88	0,01	0,02	0,09
hardcore scream	High	0,03	0,78	0,14	0,04
	Mid	0,04	0,76	0,13	0,08
	Low	0,04	0,58	0,07	0,32
black shriek	High	0,02	0,11	0,77	0,10
	Mid	0,04	0,10	0,64	0,22
death growl	Mid	0,05	0,12	0,21	0,62
	Low	0,06	0,11	0,06	0,77

Figure 3.21 – Matrice de confusion détaillée pour chacune des catégories et chacun des registres.

3.4.3 Résultats obtenus pour les techniques qui ne sont pas prises en compte par l'algorithme

Comme spécifié dans la section 3.1, la catégorie *grind inhale* (GI) et les effets *pig squeal* (PS), *deep gutturals* (DeG) et *tunnel throat* (TT) comportent trop peu d'échantillons pour être considérés comme des classes. Ces techniques peuvent cependant être testées par l'algorithme développé, afin de vérifier dans quelle classe elles pourraient être distribuées. La figure 3.22 montre la proportion de ces techniques ayant été catégorisées en voix claire (CV), *hardcore scream* (HS), *black shriek* (BS) ou *death growl* (DG).

	CV	HS	BS	DG
GI	0,11	0,27	0,15	0,47
PS	0,01	0,40	0,18	0,41
DeG	0,02	0,11	0,01	0,86
TT	0,06	0,06	0,04	0,85

Figure 3.22 – Répartition des catégories et effets n'ayant pas pu être pris en compte par l'algorithme.

Les techniques *deep gutturals* et *tunnel throat* sont étiquetées à 86% et 85% dans la

classe *death growl*. Ceci est cohérent, car ces techniques sont inspirées du *death growl*. En revanche, il n'est pas très cohérent que les effets de *grind inhale* et *pig squeal* se retrouvent en grande partie dans la catégorie *death growl*, puisqu'à l'écoute, ces techniques sont plus proches du *hardcore scream* que du *death growl*.

3.4.4 Résultats par voyelles

La figure 3.23 montre que les voyelles [a] et [i] obtiennent de bonnes précisions pour chacune des classes. En revanche, la voyelle [u] prononcée avec une technique de *black shriek* obtient un score de précision de seulement 41%. 48% des *black shriek* ont notamment été confondus avec un *death growl*. A l'inverse, le *death growl* pour la voyelle [u] est correctement étiqueté à 86%, contre 74% pour le [i] et 70% pour le [a]. Cette particularité rejoint les remarques de Hainaut (2020) sur les différences de prononciation dans les voyelles entre le chant *black* et le chant *death*. La voyelle [u] étant plus facile à produire avec la technique de *death growl* et plus difficile avec la technique de *black shriek*, il paraît cohérent qu'ils obtiennent respectivement de meilleurs et de moins bons résultats.

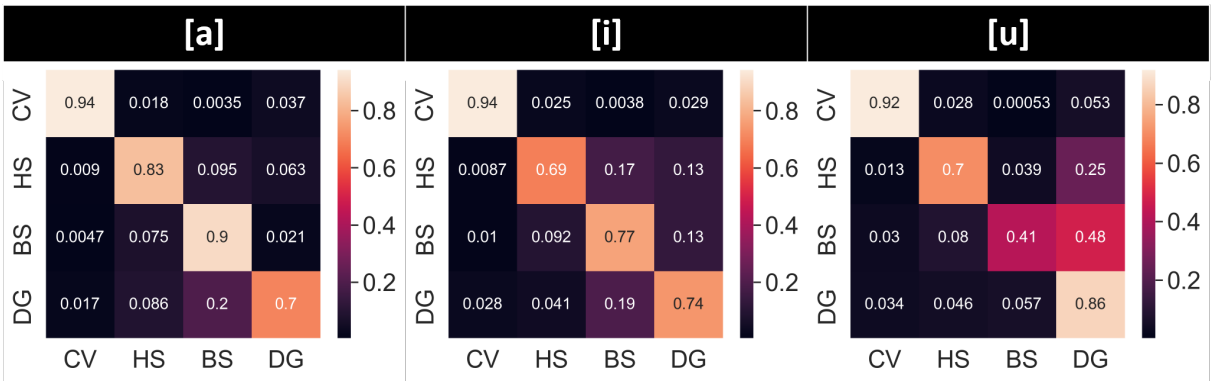

Figure 3.23 – Matrice de confusion pour chaque voyelle.

3.4.5 Résultats par notes issues du tableau d'auto-évaluation

La figure 3.24 montre que les notes du tableau d'auto-évaluation ont très peu d'influence sur la précision. Les sons associés aux notes de 1/5 possèdent même une meilleure précision que ceux associés aux notes de 5/5. Autrement dit, il semblerait que l'algorithme fonctionne aussi bien sur des sujets inexpérimentés que sur des sujets employant ces techniques quotidiennement. Il se pourrait également que le système de notation employé, imposant aux chanteurs de s'évaluer sur un degré de pratique et non sur leur performance, ne soit pas adapté à cette étude.

	1/5	2/5	3/5	4/5	5/5
précision	77,1%	74,5%	76,8%	79,6%	71,2%

Figure 3.24 – Précision pour chacune des notes issues du tableau d'auto-évaluation.

3.4.6 Résultats avec et sans tri

Les résultats avec et sans tri effectué en section 3.1 ont été comparés. En retirant ce tri, la précision baisse de 75,1% à 68,4%. Ce tri permet donc bien d'améliorer globalement les résultats de l'algorithme.

3.4.7 Résultats par sujets

Les résultats de l'algorithme sont très variables en fonction des sujets. Les sujets 10 et 17 ont par exemple obtenu un score de précision de plus de 90% dans chacune des catégories, tandis que le sujet 4 a des scores de précision parfois en dessous de 60% (voir figures 3.26, 3.27 et 3.28). Le tri effectué en section 3.1, même si il permet globalement d'augmenter la précision, n'a pas permis d'améliorer la qualité de la classification pour certains sujets, comme le sujet 3 (voir figure 3.25). 83% des *black shriek* du sujet 3 ont été étiquetées dans la catégorie *hardcore scream*. Après réécoute des enregistrements, ces *black shriek* auraient effectivement dû être retirés de la base de données, car trop proches de la technique *hardcore scream*. En revanche, pour chaque sujet, lorsqu'un tri a été effectué, cela a toujours permis d'augmenter la précision, comme il est possible de le voir pour les sujets 3, 4, 10, et 17. Il est à noter ici que ces quatre sujets sont des hommes.

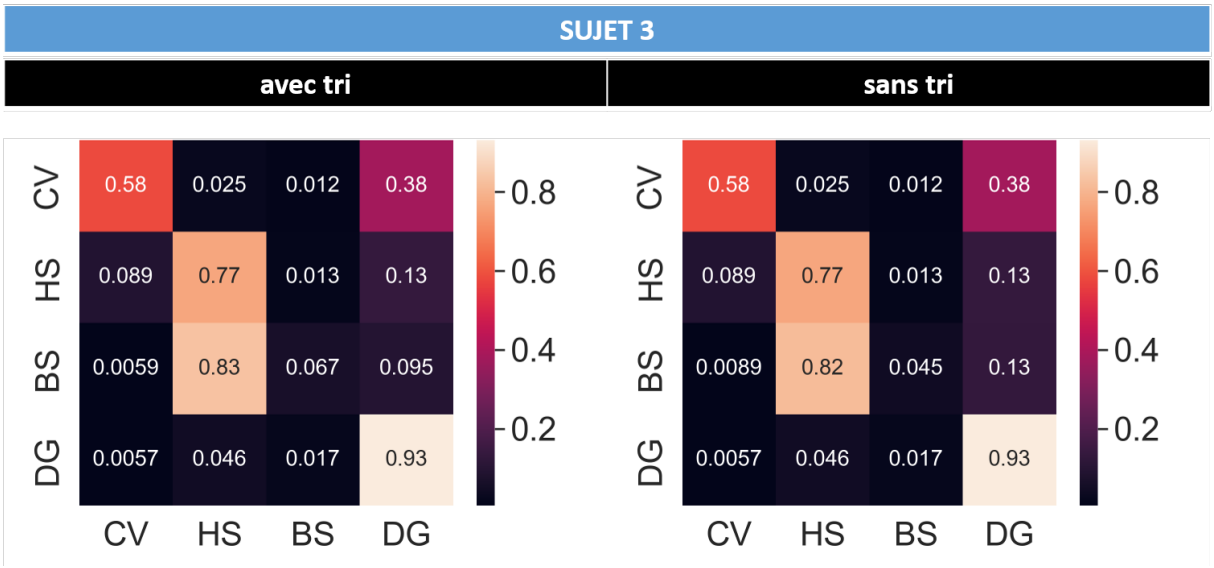

Figure 3.25 – Matrice de confusion du sujet 3 avec et sans le tri effectué en section 3.1.

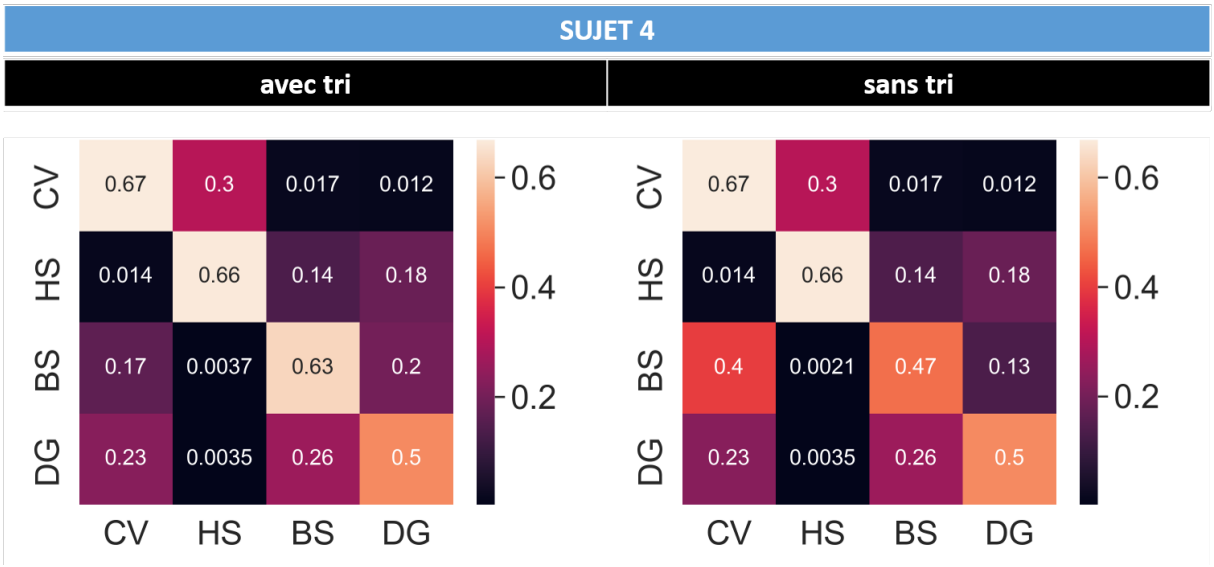

Figure 3.26 – Matrice de confusion du sujet 4 avec et sans le tri effectué en section 3.1.

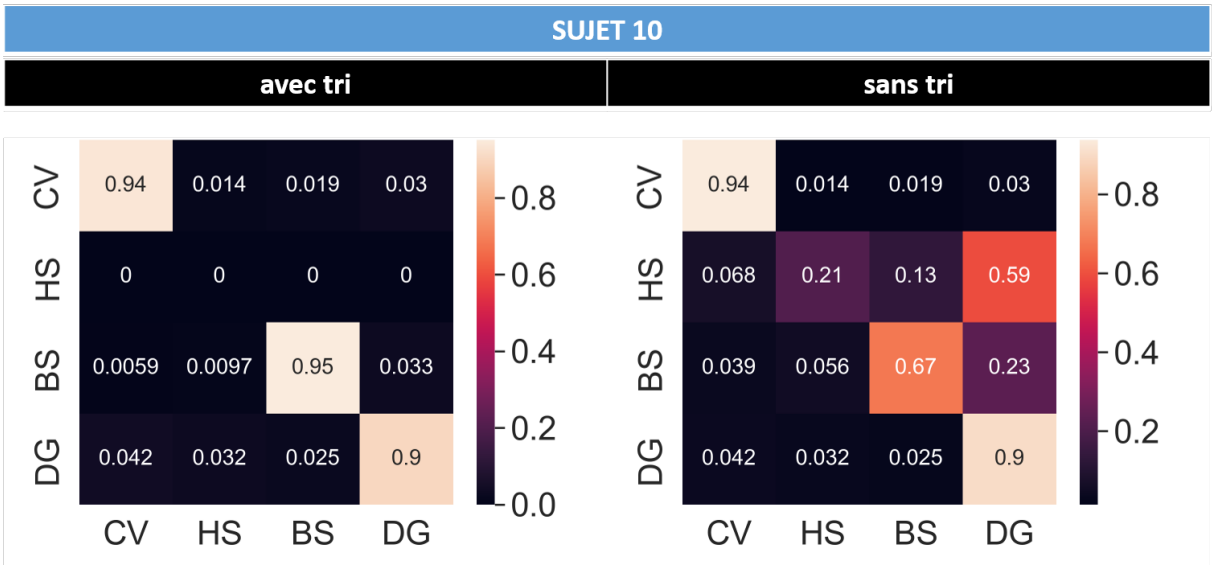

Figure 3.27 – Matrice de confusion du sujet 10 avec et sans le tri effectué en section 3.1.

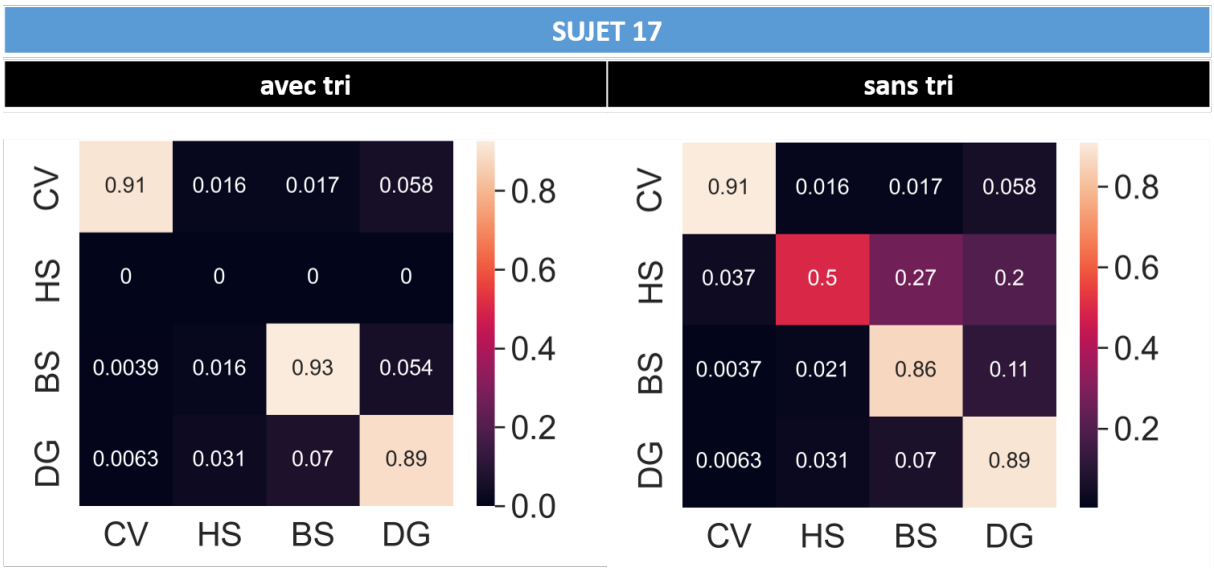

Figure 3.28 – Matrice de confusion du sujet 4 avec et sans le tri effectué en section 3.1.

3.4.8 Résultats par genre

Les femmes ont obtenu un score de précision plus élevé que les hommes pour le *hardcore scream* mais bien plus faible pour le *black shriek* et le *death growl*. Chez les femmes, 51% des *death growl* sont confondus avec des *black shriek* et 29% des *black shriek* sont confondus avec des *hardcore scream*. Ces résultats pourraient suggérer qu'il serait éventuellement intéressant de créer un modèle de classification différent, spécifiquement développé pour les voix féminines. Ces résultats doivent néanmoins être considérés avec précaution, puisqu'il existe une grande variance entre les résultats des sujets, et que les femmes ne représentent que 14,8% du total des sujets.

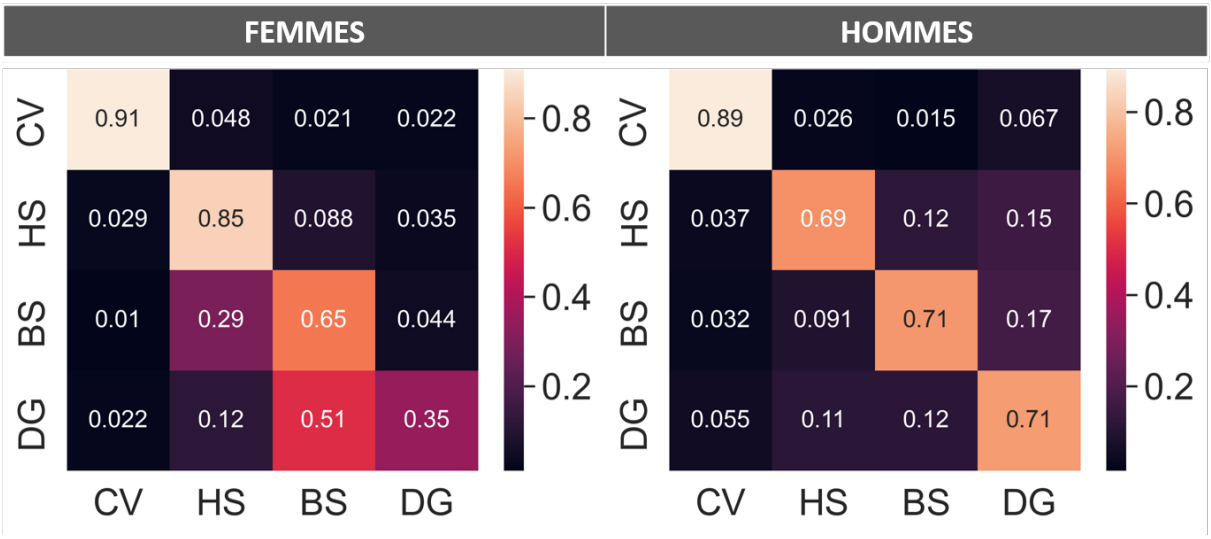

Figure 3.29 – Matrice de confusion des femmes et des hommes.

3.5 Modèles proposés pour chaque sous-genre extrême du *heavy metal*

En pratique, il est très rare qu'un chanteur soit amené à utiliser dans une même chanson tous les registres de toutes les techniques étudiées précédemment. Par exemple, le *hardcore scream* dans le registre *low* et le *death growl* dans le registre *low* possèdent très souvent la même fonction musicale, et très peu de chanteurs utilisent les deux techniques dans ce registre au sein d'une même chanson. De nouveaux modèles sont donc créés, chaque modèle étant ainsi associé à un des sous-genres extrêmes du *heavy metal* (voir figure 1.1). L'entraînement des modèles se fait donc avec des catégories et des registres pertinents pour le genre étudié.

3.5.1 Modèle *death metal*

Comme montré en figure 1.1, les chanteurs de *death metal* utilisent principalement des techniques de *death growl* et de *black shriek*. La matrice de confusion du modèle *death metal* est présentée en figure 3.30. Ce modèle obtient un score de précision de 82,8%.

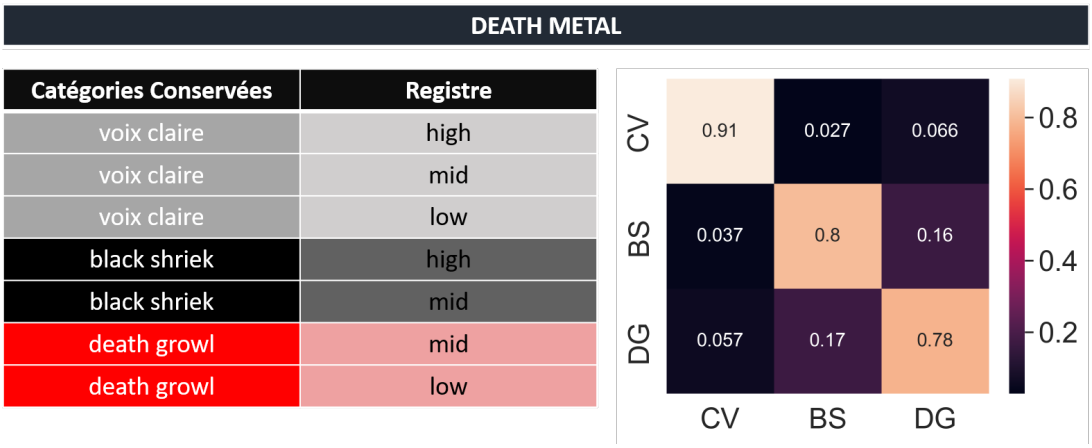

Figure 3.30 – Matrice de confusion du modèle *death metal*.

3.5.2 Modèle *black metal*

Comme montré en figure 1.1, les chanteurs de *black metal* utilisent principalement des techniques de *black shriek*, et de *hardcore scream* dans les registres *high* et *mid*. La matrice de confusion de ce modèle est présentée en figure 3.31. Ce modèle obtient un score de précision de 84,8%.

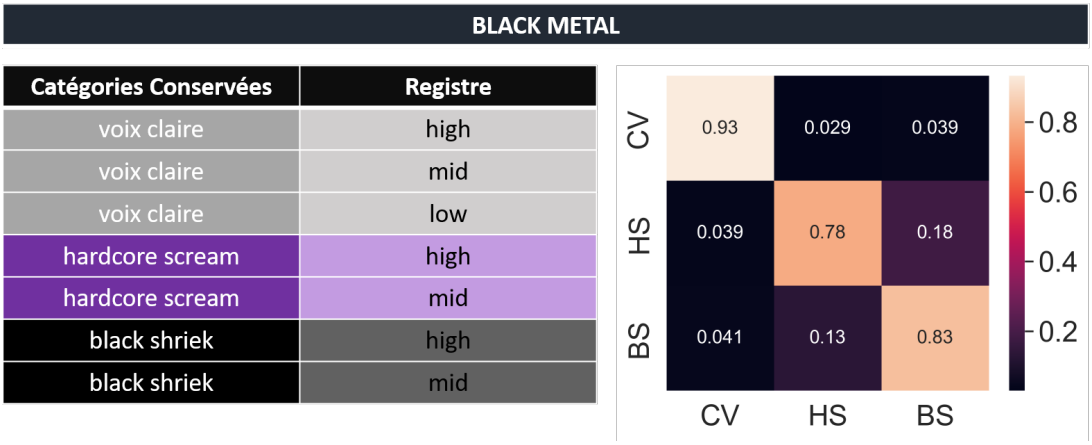

Figure 3.31 – Matrice de confusion du modèle *death metal*.

3.5.3 Modèle *metalcore*

Comme montré en figure 1.1, les chanteurs de *metalcore* utilisent principalement des techniques de *hardcore scream*. La matrice de confusion de ce modèle est présentée en figure 3.32. Ce modèle obtient un score de précision de 96,4%.

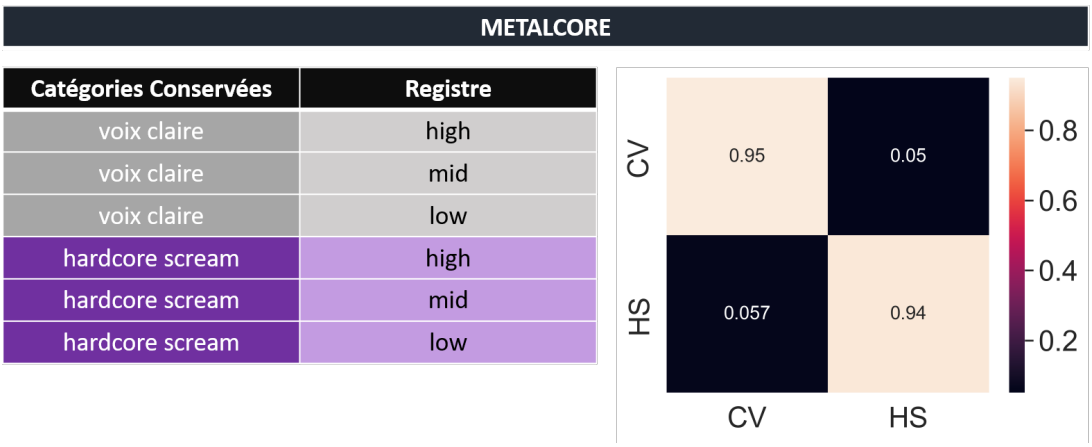

Figure 3.32 – Matrice de confusion du modèle *metalcore*.

3.5.4 Modèle *grindcore* (et *deathcore*)

Comme montré en figure 1.1, les chanteurs de *grindcore* et de *deathcore* utilisent toutes les techniques de saturation extrême. Le *deathcore* est tout simplement un sous-genre issu de la fusion de *death metal* et du *metalcore*. Le *grindinhale*, avec ce modèle, est détecté à 62,9% dans la catégorie *death growl*. Ce modèle obtient un score de précision de 77,9%.

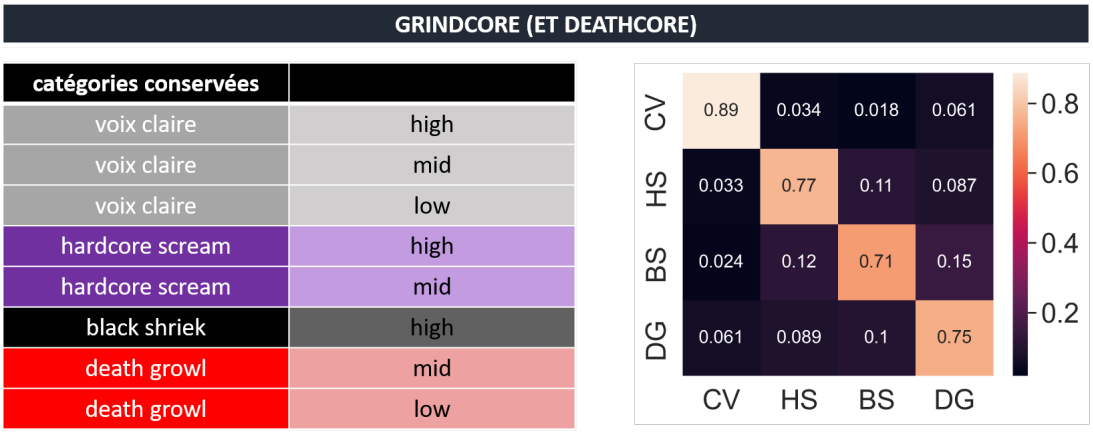

Figure 3.33 – Matrice de confusion du modèle *grindcore* (et *deathcore*).

3.6 Conclusion pour la partie analyse des résultats

Les meilleures précisions sont obtenues en utilisant les 18 premiers DAFCC (avec 32 filtres) et en entraînant un modèle de perceptron multicouche avec 200 neurones pour chacune des deux sous-couches.

Les analyses détaillées ont montré que la confusion devient particulièrement critique lorsque les registres de chants saturés se recoupent, par exemple en *mid* entre le *death growl* et le *black shriek* ou en *low* entre le *hardcore scream* et le *death growl*. Pour remédier à ce problème, de nouveaux modèles ont été conçus, des modèles adaptés au genre de musique *metal* ciblé et aux techniques vocales qui y sont produites. Le modèle *death metal* a obtenu un score de 82,8%, le *black metal* un score de 84,8%, le *metalcore* un score de 94,6% et le *deathcore* un score de 77,9%.

L'analyse par sujets a montré une grande variabilité dans les résultats en précision. De plus, les femmes obtiennent de moins bons résultats que les hommes pour les techniques de *black shriek* et de *death growl*, mais de meilleurs résultats pour la technique de *hardcore scream*. Ces résultats pourraient suggérer qu'il faudrait utiliser des modèles différents pour les hommes et pour les femmes.

Le tri des données et l'*undersampling* aléatoire permettent d'améliorer significativement la précision de l'algorithme. L'analyse des résultats par sujet a néanmoins montré que

le tri des données pouvait être amélioré.

Conclusion

Ce mémoire se présente comme une étude préliminaire, en vue du développement d'un plug-in qui permettrait de détecter en temps réel les différentes techniques de distorsion vocale utilisées dans le *metal* extrême, et de rediriger le signal vers des bus de traitements adaptés. Une nouvelle taxonomie des techniques les plus extrêmes a ainsi été conçue à partir des taxonomies existantes, et des différentes remarques des chanteurs.

Suite à la création de cette taxonomie, une base de données comportant les voix de 27 chanteurs produisant ces techniques pour plusieurs voyelles et pour un texte de leur choix a été conçue. Afin d'avoir un modèle potentiellement applicable en temps réel, les sons de la base de données ont été découpés en trames de 1024 échantillons.

Plusieurs descripteurs ont été testés, à savoir les MFCC, le contraste spectral, et les DAFCC. Ce sont les DAFCC qui donnent la meilleure précision. Les DAFCC (Data Adjusted Frequencies Cepstral Coefficients) sont de nouveaux descripteurs dont le calcul est inspiré de celui des MFCC, et élaborés lors de ce mémoire, afin d'adapter le découpage fréquentiel aux données et à leur répartition dans les différentes classes. Plusieurs modèles de *Machine Learning* ont également été comparés : la forêt d'arbres décisionnels, le perceptron multicouche et la classification naïve bayésienne. Le perceptron multicouche est le modèle montrant le plus de potentiel pour cette approche. En extrayant les 18 premiers DAFCC à partir de 32 filtres, et en entraînant un perceptron multicouche comportant 200 neurones pour chacune de ses deux couches intermédiaires, la précision atteint 75,2%. Ce résultat peut être amélioré en créant des modèles spécifiques aux sous-genres extrêmes du *heavy metal*. Le modèle *death metal* a obtenu un score de précision de 82,8%, le *black metal* un score de 84,8%, le *metalcore* un score de 94,6% et le *grindcore* (et *deathcore*) un score de 77,9%.

Ces résultats, bien que satisfaisants, sont cependant très variables en fonction des chanteurs. Il pourrait donc être intéressant de proposer un module de calibration, afin que les chanteurs puissent adapter le modèle à leur timbre avant son utilisation. Les résultats assez différents obtenus pour les femmes et pour les hommes pourraient également suggérer qu'il faudrait concevoir deux modèles différents, un pour chacun des genres. Cependant, le nombre de femmes enregistrées et la variabilité des résultats par sujet ne permettent pas d'élaborer une conclusion définitive à ce propos. Il pourrait donc être intéressant, dans une perspective à court terme, d'enregistrer plus de voix féminines pour pouvoir en déduire si une telle solution peut s'avérer pertinente.

Un certain nombre d'extraits sonores de la base de données ont été retirés de celle-

ci, car ils ont été considérés comme trop peu représentatifs de la catégorie qu'ils devaient décrire. Ce tri a permis d'améliorer les performances globales de l'algorithme, mais reste perfectible. Il serait intéressant dans une approche future de soumettre cette base de données à un panel d'experts (chanteurs et ingénieurs du son spécialisés dans le *metal*), afin qu'ils classent eux-mêmes ces données dans les différentes catégories.

Ces différents résultats ont montré un grand potentiel d'application en temps réel, en partie dû au faible temps de calcul des DAFCC. Suite à cette approche préliminaire, il reste donc à aborder cette perspective et à développer le plug-in en conséquence.

Les DAFCC se sont avérés particulièrement performants pour ce problème de *Machine Learning*. Il pourrait être intéressant d'appliquer cette méthode de calcul à plusieurs problèmes de classification supervisée d'extraits audio afin notamment de comparer les performances de la classification sur la base des DAFCC avec celles de la classification utilisant les MFCC.

Bibliographie

Alsouda, Y., Pllana, S. et Kurti, A. (2018). A machine learning driven IoT solution for noise classification in smart cities. *arXiv preprint arXiv :1809.00238*.

Aly, M. (2005). Survey on multiclass classification methods. *Neural Netw*, 19(1-9):9.

Azarloo, A. et Farokhi, F. (2012). Automatic musical instrument recognition using K-NN and MLP neural networks. *In 2012 Fourth International Conference on Computational Intelligence, Communication Systems and Networks*, pages 289–294. IEEE.

Bailly, L. (2009). *Interaction entre cordes vocales et bandes ventriculaires en phonation : exploration in-vivo, modélisation physique, validation in-vitro.* PhD Thesis, Université du Maine, Le Man.

Bonada, J. et Blaauw, M. (2013). Generation of growl-type voice qualities by spectral morphing. *In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6910–6914. IEEE.

Breiman, L., Friedman, J. H., Olshen, R. A. et Stone, C. J. (1984). *Classification and regression trees*. Chapman & Hall.

Camastra, F. et Vinciarelli, A. (2008). *Machine learning for audio, image and video analysis : theory and applications*. Springer, Londres.

Caruana, R. et Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *In Proceedings of the 23rd international conference on Machine learning*, pages 161–168.

Chevaillier, G., Guilbault, R., Renard, J.-N., Herman, P. et Tran Ba Huy, P. (2011). La voix «saturée» du chanteur rock métal, un mécanisme supraglottique performant. *La polyclinique de phoniatrie Dr Chevaillier*.

Dal Pozzolo, A., Caelen, O. et Bontempi, G. (2015). When is undersampling effective in unbalanced classification tasks ? *In Joint european conference on machine learning and knowledge discovery in databases*, pages 200–215. Springer.

Davis, S. et Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366.

Fernández, A., Garcia, S., Herrera, F. et Chawla, N. V. (2018). SMOTE for learning from imbalanced data : progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905.

Fitch, W. T., Neubauer, J. et Herzel, H. (2002). Calls out of chaos : the adaptive significance of nonlinear phenomena in mammalian vocal production. *Animal behaviour*, 63(3):407–418.

Ganchev, T., Fakotakis, N. et Kokkinakis, G. (2005). Comparative evaluation of various MFCC implementations on the speaker verification task. *In Proceedings of the SPECOM*, volume 1, pages 191–194.

Gentilucci, M., Ardaillon, L. et Liuni, M. (2019). Composing vocal distortion : A tool for real-time generation of roughness. *Computer Music Journal*, 42(4):26–40.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow : Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, Inc., Sébastopol, CA, USA.

Hainaut, B. (2020). Des vocalités «bestiales» ? Caractériser les voix bruitées du black metal. *Volume !*, 16(1):145–161.

He, H., Bai, Y., Garcia, E. A. et Li, S. (2008). ADASYN : Adaptive synthetic sampling approach for imbalanced learning. *In 2008 IEEE international joint conference on neural networks*, pages 1322–1328. IEEE.

Hoerl, A. E. et Kennard, R. W. (1970). Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Iheme, L. O. et Ozan, (2019). Multiclass digital audio segmentation with MFCC features using naive Bayes and SVM classifiers. *In 2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5. IEEE.

Jiang, D.-N., Lu, L., Zhang, H.-J., Tao, J.-H. et Cai, L.-H. (2002). Music type classification by spectral contrast feature. *In Proceedings. IEEE International Conference on Multimedia and Expo*, volume 1, pages 113–116. IEEE.

Jones, T. M., Trabold, M., Plante, F., Cheetham, B. M. G. et Earis, J. E. (2001). Objective assessment of hoarseness by measuring jitter. *Clinical Otolaryngology & Allied Sciences*, 26(1):29–32.

Kent, R. D. et Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths : A review. *Journal of communication disorders*, 74:74–97.

Kingma, D. P. et Ba, J. (2015). Adam : A method for stochastic optimization. *In International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Lin, W.-C., Tsai, C.-F., Hu, Y.-H. et Jhang, J.-S. (2017). Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409:17–26.

Loscos, A. et Bonada, J. (2004). Emulating rough and growl voice in spectral domain. *In Proceedings of the International Conference on Digital Audio Effects*, pages 49–52, Naples, Italy. Citeseer.

Mandal, P., Nath, I., Gupta, N., Jha, M. K., Ganguly, D. G. et Pal, S. (2020). Automatic music genre detection using artificial neural networks. *In Intelligent Computing in Engineering*, pages 17–24, Singapour. Springer.

McGlashan, J., Sadolin, C. et Kjelin, H. (2007). Can vocal effects such distortion, growling, rattle and grunting be produced without traumatising the vocal folds. *In Proc. of the Pan European Voice Conference. Groningen, The Netherlands*, Groningue, Pays-Bas.

McGlashan, J., Sayles, M., Sadolin, C. et Kjelin, H. (2013). Vocal effects in singing : a study of intentional distortion using laryngostroboscopy and electrolaryngography. *In 10th International Conference on Advance in Quantitative Laryngology, Voice and Speech Research*.

McGlashan, J., Thuesen, M. A. et Sadolin, C. (2017). Overdrive and edge as refiners of "belting" ? : an empirical study qualifying and categorizing "belting" based on audio perception, laryngostroboscopic imaging, acoustics, LTAS, and EGG. *Journal of Voice*, 31(3):385–e11.

Medina, Y. O., Beltrán, J. R. et Baldassarri, S. (2020). Emotional classification of music using neural networks with the MediaEval dataset. *Personal and Ubiquitous Computing*, pages 1–13.

Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Vincent, E., Raj, B. et Virtanen, T. (2017). DCASE 2017 challenge setup : Tasks, datasets and baseline system. *In DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*.

Myint, E. S. et Ni, N. (2020). *Audio Classification in Speech and Music by using Neural Network : Multilayer Perceptron*. PhD Thesis, University of Computer Studies, Yangon, Birmanie.

Netter, F. H. et Scott, J. (2007). *Atlas d'anatomie humaine*. Masson.

Nieto, O. (2008). *Voice transformations for extreme vocal effects*. Mémoire de master, Universitat Pompeu Fabra, Barcelone, Espagne.

Nieto, O. (2013). Unsupervised clustering of extreme vocal effects. *In Proc. 10th Int. Conf. Advances in Quantitative Laryngology*, pages 115–116, Cincinnati, Ohio, USA.

Nwe, T. L., Shenoy, A. et Wang, Y. (2004). Singing voice detection in popular music. *In Proceedings of the 12th annual ACM international conference on Multimedia*, pages 324–327.

Oo, M. M. (2018). Comparative study of MFCC feature with different machine learning techniques in acoustic scene classification. *International Journal of Research and Engineering*, 5:439–444.

Picone, J. W. (1993). Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, 81(9):1215–1247.

Purcell, N. J. (2003). *Death metal music : The passion and politics of a subculture*. McFarland, Jefferson, NC, USA.

Raschka, S. et Mirjalili, V. (2017). *Python machine learning : Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd, UK, 2nd édition.

Rocamora, M. et Herrera, P. (2007). Comparing audio descriptors for singing voice detection in music audio files. *In Brazilian symposium on computer music, 11th*, volume 26, page 27, San Pablo, Brésil.

Sadolin, C. (2000). *Complete vocal technique*. Shout Publishing, Copenhague, Danemark.

Sakakibara, K.-I., Fuks, L., Imagawa, H. et Tayama, N. (2004). Growl voice in ethnic and pop styles. *In Proc. Int. Symp. on Musical Acoustics*, Nara, Japon. Citeseer.

Scotto Di Carlo, N. (1991). La voix chantée. *La Recherche*, XXIII(235):1016–1025.

Slaney, M. (1998). Auditory toolbox. *Interval Research Corporation, Tech. Rep*, 10(1998): 1194.

Smialek, E., Depalle, P. et Brackett, D. (2012). A spectrographic analysis of vocal techniques in extreme metal for musicological analysis. *In ICMC*, pages 88–93.

Sundberg, J. et Rossing, T. D. (1987). *The science of singing voice*. Northern Illinois University Press, DeKalb, Illinois, USA.

Thambi, S. V., Sreekumar, K. T., Kumar, C. S. et Raj, P. R. (2014). Random forest algorithm for improving the performance of speech/non-speech detection. *In 2014 First International Conference on Computational Systems and Communications (ICCSC)*, pages 28–32, Trivandrum, Inde. IEEE.

Thuesen, M. A., McGlashan, J. et Sadolin, C. (2017). Curbing—the metallic mode in-between : an empirical study qualifying and categorizing restrained sounds known as curbing based on audio perception, laryngostroboscopic imaging, acoustics, LTAS, and EGG. *Journal of Voice*, 31(5):644–e1.

Tsai, C.-G., Wang, L.-C., Wang, S.-F., Shau, Y.-W., Hsiao, T.-Y. et Auhagen, W. (2010). Aggressiveness of the growl-like timbre : Acoustic characteristics, musical implications, and biomechanical mechanisms. *Music Perception*, 27(3):209–222.

Verma, A. et Kumar, A. (2005). Introducing roughness in individuality transformation through jitter modeling and modification. *In Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-05*, pages 5–8. IEEE.

Wei, P., He, F., Li, L. et Li, J. (2020). Research on sound classification based on SVM. *Neural Computing and Applications*, 32(6):1593–1607.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D. et Povey, D. (2002). The HTK book. *Cambridge university engineering department*, 3(175):12.

Yumoto, E., Sasaki, Y. et Okamura, H. (1984). Harmonics-to-noise ratio and psychophysical measurement of the degree of hoarseness. *Journal of Speech, Language, and Hearing Research*, 27(1):2–6.

Zelkowitz, M. (2010). *Improving the Web*. Elsevier.

Zhang, Y. et LV, D.-j. (2015). Selected features for classifying environmental audio data with random forest. *The Open Automation and Control Systems Journal*, 7(1).

Discographie

Black Metal, Venom [CD-ROM]. RU, Newcastle : Neat Records. 1982.

Corporal Violence, Genocide Of Prescription [CD-ROM]. 3283164 Records DK. 2009.

Creature, Baest [CD-ROM]. Allemagne, Dortmund : Century Media Records Ltd. 2022.

Death Atlas, Cattle Decapitation [CD-ROM]. USA, Californie, Agoura Hills : Metal Blade Records. 2019.

Death By Metal, Mantas [CD-ROM]. USA, Philadelphie : Relapse Records. 1984.

Death Metal, Possessed [CD-ROM]. Pays-Bas, Zeist : VIC Records. 1984.

Diamond, Stick To Your Guns [CD-ROM]. USA, Californie, Los Angeles : Sumerian Records. 2012.

Eight Headed Serpent, Impaled Nazarene [CD-ROM]. France, Beaurainville : Osmose Productions. 2021.

INRI, Sarcofago [CD-ROM]. Brésil, Belo Horizonte : Cogumelo. 1987.

Insurrectionist (Deluxe), Dead/Awake [CD-ROM]. USA, Californie, Los Angeles : We Are Triumphant. 2021.

Necrobreed, Benighted [CD-ROM]. France, Marseille : Season Of Mist. 2017.

No Peace in Hell, Get The Shot [CD-ROM]. Angleterre, Londres : Demons Run Amok Entertainment. 2014.

Nymphetamine, Cradle Of Filth [CD-ROM]. USA, New-York : Roadrunner Records. 2004.

Purification Through Violence, Dying Fetus [CD-ROM]. USA, Upper Darby, Pennsylvanie : Relapse. 1996.

Recollections of the Insane, Schizophrenia [CD-ROM]. USA, Ohio, Cleveland : Redefining Darkness Records. 2022.

There’s Always Blood At The End Of The Road, Wiegedood [CD-ROM]. Allemagne, Dortmund : Century Media Records. 2022.

Welcome To Sludge City, Annotations of an Autopsy [CD-ROM]. USA, Calofornie, Westminster : This City Is Burning Records. 2007.

You Will Never Be One Of Us, NAILS [CD-ROM]. Allemagne, Donzdorf : Nuclear Blast. 2016.

Annexe A

Tableau d’auto-évalution et informations sur les chanteurs

Cette annexe présente le tableau d’auto-évaluation des chanteurs, et les différentes informations ayant été récupérées lors de la session d’enregistrement : le genre du chanteur ou de la chanteuse (M pour masculin, F pour féminin), si l’enregistrement a été effectué en direct ou à distance, la distance au microphone, le microphone utilisé, la carte son utilisée et le logiciel utilisé pour la captation.

Alias	Genre	Enregistrement	Dist Au Micro (cm)	Micro	Carte Son	DAO
Sujet 1	M	Direct	10	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 2	M	Direct	3	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 3	M	Distance	10	Shure Beta 58a	Zoom R16	Cubase
Sujet 4	M	Direct	5	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 5	M	Direct	5	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 6	F	Direct	2	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 7	M	Direct	2	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 8	M	Distance	2	SM58	Focusrite Scarlett Solo	Reaper
Sujet 9	M	Direct	2	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 10	M	Direct	10	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 11	M	Direct	2	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 12	M	Direct	5	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 13	M	DIstance	5	SM58	Yamaha N12	ProTools
Sujet 14	M	Direct	3	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 15	M	Direct	5	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 16	M	Direct	3	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 17	M	Distance	5	SM58	Focusrite Scarlett 6i6	Studio One
Sujet 18	F	Distance	2	SM58	RME babiface pro	Reaper
Sujet 19	M	Direct	2	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 20	F	Direct	3	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 21	M	Direct	3	SM58	Focusrite Scarlett 2i2	ProTools
Sujet 22	M	Direct	1	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 23	M	Direct	5	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 24	M	Direct	5	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 25	F	Direct	2	SM58	Scarlett Solo	ProTools
Sujet 26	M	Direct	5	SM58	Focusrite Scarlett 6i6	ProTools
Sujet 27	M.	Direct	2	SM58	Focusrite Scarlett 6i6	Reaper

Figure A.1 – Informations sur les chanteurs.

Alias	VoixClaire_High	VoixClaire_Mid	VoixClaire_Low	BlackShriek_High	BlackShriek_Mid
Sujet 1	1	1	1	0	3
Sujet 2	3	3	3	4	3
Sujet 3	1	1	1	3	4
Sujet 4	5	5	5	5	5
Sujet 5	4	4	4	4	5
Sujet 6	5	5	5	0	4
Sujet 7	2	4	4	2	3
Sujet 8	4	3	2	5	5
Sujet 9	2	3	2	4	4
Sujet 10	2	2	1	4	5
Sujet 11	5	5	5	2	2
Sujet 12	5	5	5	5	5
Sujet 13	0	0	0	4	0
Sujet 14	3	4	2	5	5
Sujet 15	2	2	2	3	3
Sujet 16	4	3	2	0	1
Sujet 17	2	2	2	4	4
Sujet 18	5	5	5	3	2
Sujet 19	1	1	1	4	3
Sujet 20	3	3	3	2	2
Sujet 21	0	0	0	3	3
Sujet 22	3	2	2	2	1
Sujet 23	4	4	4	1	1
Sujet 24	5	5	5	0	0
Sujet 25	1	1	1	3	3
Sujet 26	2	2	2	2	2
Sujet 27	4	4	4	1	1

Figure A.2 – Tableau d’autoévaluation (partie 1).

Alias	DeathGrowl_Mid	DeathGrowl_Low	HardcoreScream_High	HardcoreScream_Mid	HardcoreScream_Low
Sujet 1	5	5	0	3	0
Sujet 2	5	5	0	1	3
Sujet 3	5	5	3	0	0
Sujet 4	5	5	5	5	5
Sujet 5	5	5	0	0	0
Sujet 6	5	4	0	0	0
Sujet 7	4	4	0	3	3
Sujet 8	5	5	4	4	3
Sujet 9	4	3	1	2	3
Sujet 10	5	4	3	3	3
Sujet 11	3	3	4	4	4
Sujet 12	5	5	5	5	5
Sujet 13	4	5	0	0	0
Sujet 14	5	5	0	0	0
Sujet 15	2	2	4	5	4
Sujet 16	1	1	4	4	3
Sujet 17	5	5	2	3	3
Sujet 18	2	2	2	3	3
Sujet 19	1	1	5	5	4
Sujet 20	4	4	4	5	4
Sujet 21	4	4	0	0	0
Sujet 22	2	2	5	3	4
Sujet 23	1	3	5	4	2
Sujet 24	0	0	3	0	0
Sujet 25	3	5	1	1	0
Sujet 26	5	4	1	1	1
Sujet 27	1	1	4	4	4

Figure A.3 – Tableau d’autoévaluation (partie 2).

Alias	GrindInhale	PigSqueal	DeepGutturals	TunnelThroat
Sujet 1	0	0	4	0
Sujet 2	0	0	0	5
Sujet 3	0	4	1	1
Sujet 4	5	5	0	5
Sujet 5	0	0	0	0
Sujet 6	0	0	0	0
Sujet 7	0	0	2	0
Sujet 8	0	2	5	3
Sujet 9	3	1	1	1
Sujet 10	0	0	0	3
Sujet 11	1	2	3	3
Sujet 12	3	3	3	1
Sujet 13	0	0	0	0
Sujet 14	0	0	4	4
Sujet 15	0	0	0	0
Sujet 16	2	1	1	1
Sujet 17	3	2	5	2
Sujet 18	0	3	2	2
Sujet 19	0	0	0	0
Sujet 20	0	0	0	0
Sujet 21	0	0	4	0
Sujet 22	0	2	3	2
Sujet 23	0	0	2	0
Sujet 24	0	0	0	0
Sujet 25	0	5	5	5
Sujet 26	1	0	3	1
Sujet 27	0	1	0	1

Figure A.4 – Tableau d’autoévaluation (partie 3).

Annexe B

Exemples proposés aux chanteurs lors des enregistrements

Cette annexe présente les différents exemples ayant été diffusés dans le casque des chanteurs avant chaque enregistrement.

Catégorie	Registre	Exemple
voix claire	high	\
	mid	\
	low	\
hardcore scream	high	Get The Shot – Cold Hearted
	mid	Stick To Your Guns – Against Them All
	low	NAILS – You Will Never Be One Of Us
black shriek	high	Dimmu Borgir – Sorgens Kammer Del II
	mid	Marduk – Panzer Division Marduk
death growl	mid	Cannibal Corpse – Evisceration Plague
	low	Six Feet Under – Seed Of Filth

Figure B.1 – Exemples musicaux proposés aux chanteurs pour chaque catégorie.

Effet	Exemple	TimeCode
pig squeal	Genocide Of Prescription – The Man Who Build A God	de 01:00 à 01:06
deep gutturals	Dead/Awake – The Pale Horse 2.0	de 02:53 à 02:59
tunnel throat	Dimmu Borgir – Sorgens Kammer Del II	de 00:00 à 00:08

Figure B.2 – Exemples musicaux proposés aux chanteurs pour chaque effet.