Amaury BODIN,
Vincent BERNARD

Groupe A4

Rapport Projet E4 : Vocodeur de phase

Table des matières

1 Introduction

Nous allons réaliser un projet qui nous permettra d'étudier des signaux vocaux. Nous verrons comment il est possible de jouer avec les pitchs pour les réutiliser dans le cinéma par exemple. Pour cela nous allons trouver un moyen de jouer avec la vitesse de parole sans modifier le pitch. Dans un second temps, nous allons trouver un moyen de modifier un pitch sans avoir à modifier la vitesse de la voix. Pour terminer, nous devons trouver le moyen de robotiser une voix.
Pour ce faire nous allons utiliser le logiciel Matlab ainsi que plusieurs fonctions qui nous ont été fournies pour réaliser le programme principale Vocodeur.m. Nous allons travailler sur des audios qui nous ont également été donnés.

2 Partie A : Modification du pitch, modification de la vitesse

2.1 Pitch initial (sans modification) :

Dans un 1er temps on décide d'écouter les différents audios sur lesquels nous allons travailler et nous observerons également les différentes courbes dans les domaines temporels, fréquentiels et le spectrogramme qui leur sont caractéristiques.

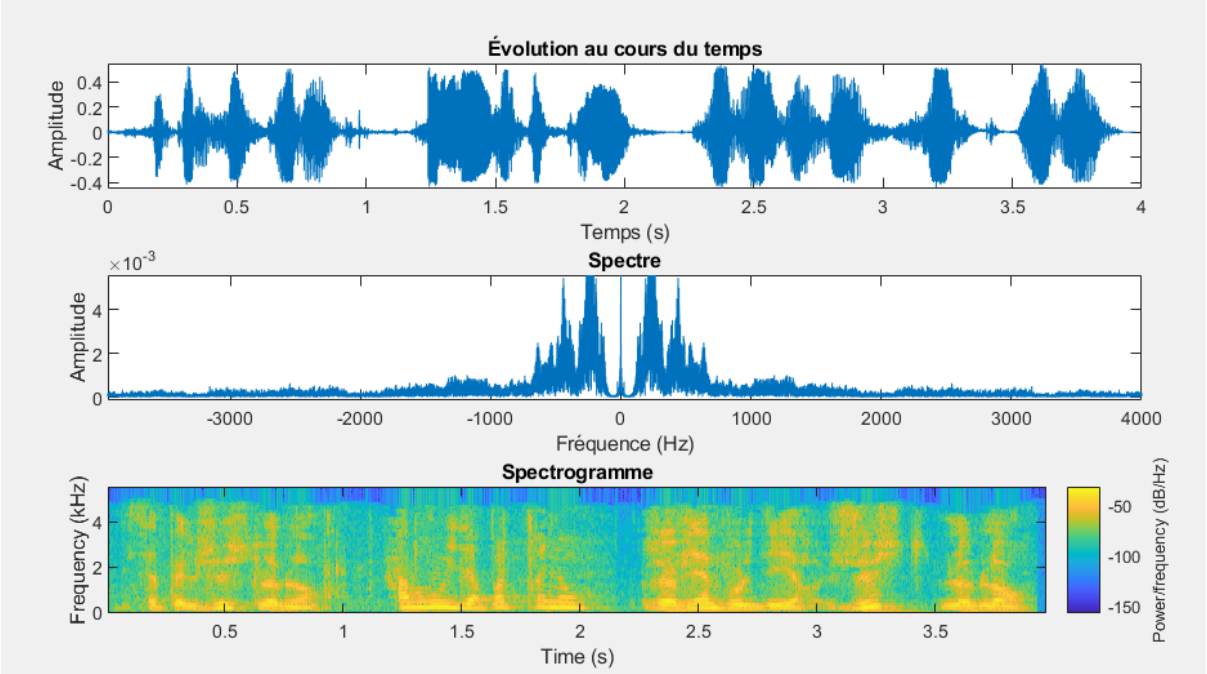

Fig 1 : Représentation du signal Dinner.wav dans les domaines temporels, fréquentiels et le Spectrogramme

Grâce à l'affichage de l'évolution du son au cours du temps, on pourra constater que la durée de l'audio fait 4 secondes. En s'appuyant sur le graphique nous pouvons visualiser les silences et les pauses dans une discussion lorsque l'amplitude du signal est à 0. On peut donc également visualiser l'intensité et l'amplitude du signal. En effet, un signal plus fort aura une amplitude plus élevée.

Si on s'intéresse au spectre du signal sonore, on peut également constater la/les fréquences des différentes voix présentes dans un audio. Ici, tous les grands pics correspondants à des voix sont situés entre 100 et 600 Hz. En effet, ces fréquences correspondent aux voix d'hommes et de femmes qui parlent. Cette représentation permet d'avoir une représentation de l'ensemble des fréquences disponibles dans l'audio mais ne permet pas de distinguer les moments où elles sont utilisées dans l'audio.

Le spectrogramme lui nous permet d'avoir une représentation visuelle du spectre fréquentiel du signal audio en fonction du temps. Il peut donc nous aider à identifier les différentes composantes fréquentielles présentes dans un audio, et à quel moment elles sont présentes dans l'audio. On peut donc constater des changements brusques dans le son.

2.2 Modification de la vitesse

La voix n'est pas un signal stationnaire c'est pourquoi il est parfois très difficile de l'étudier et d'effectuer des changements sur un signal vocal. En effet, un signal stationnaire possède des caractéristiques tels que la moyenne, la variance, et la densité spectrale de puissance restent constantes tout au long du signal. La voix humaine ne peut naturellement pas satisfaire ces conditions, la fondamentale de la voix peut changer, le débit, l'intensité, le timbre de la voix peuvent varier tout au long d'une conversation. C'est pourquoi on ne peut pas considérer la voix comme un signal stationnaire. On doit donc trouver des techniques pour pallier cela, nous allons donc considérer que la voix peut être considérée comme un signal stationnaire sur une très courte période de 20 à 30 ms.

Pour modifier la vitesse d'un enregistrement, la ralentir ou l'accélérer, on doit tout d'abord définir le taux de ralentissement ou d'accélération de l'audio, on doit donc le choisir inférieur à 1 pour le ralentir et supérieur à 1 pour l'accélérer.
Nous allons ensuite appeler une fonction nommée PVoc qui va effectuer plusieurs tâches. Cette fonction prend en paramètres :

1. Le signal que nous voulons modifier (x)
2. Le rapport entre la vitesse d'origine et la vitesse d'arrivée (rapp)
3. Le nombre de points sur lesquels la transformation de Fourier rapide (FFT) va s'effectuer (Nfft). Cela détermine la résolution fréquentielle de la transformation.
4. La longueur de la fenêtre de pondération de la transformation de Fourier à court terme (Nwind). Cette fenêtre de pondération est la fenêtre de Hanning.

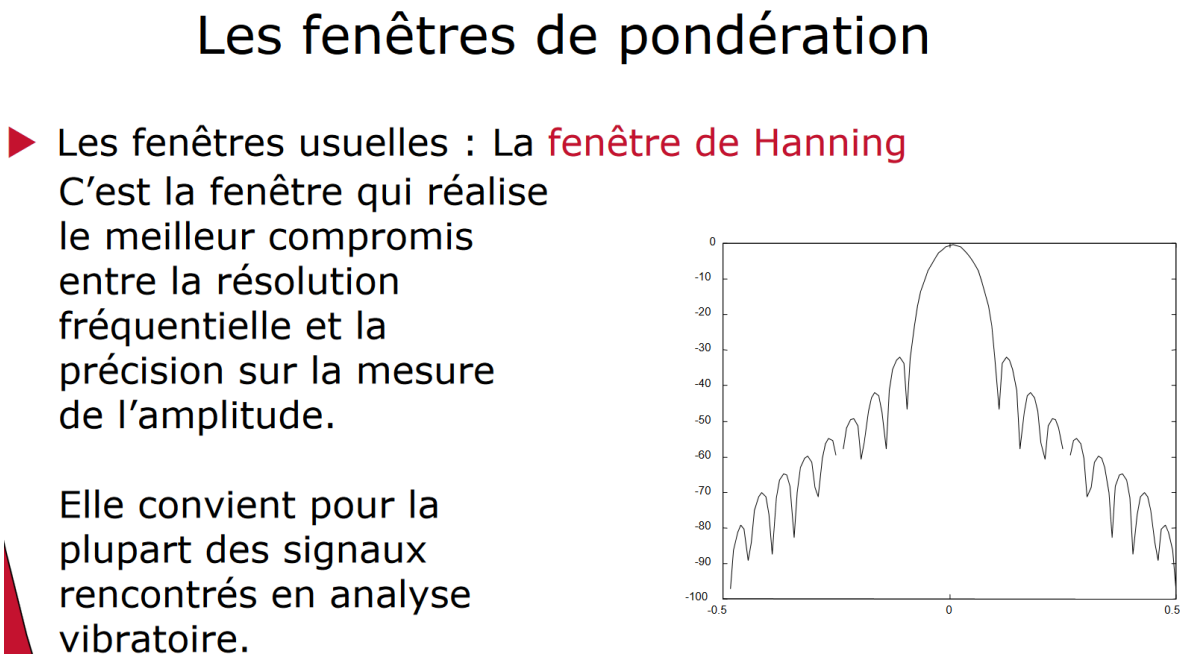

Fig 2 : Représentation de la fenêtre de Hanning

Si les paramètres 3) ou 4) ne sont pas renseignés, la fonction Pvoc utilise des paramètres par défaut. Pour 3) Nfft=1024, et pour 4) Nwind=Nfft. Cela permet de ne pas avoir d'erreurs lors de l'exécution.

Pour accélérer un son, un dialogue sans modification du son, nous préparons l'appel de la fonction TFCT (Transformée de Fourier à Court Terme).
Nous notons 'Nov' le nombre de points de recouvrement des fenêtres qui est le quart de 'Nfft', ce qui signifie que chaque fenêtre de la TFCT se chevauchera de 25%. On définit également un facteur d'échelle 'scf' qui permet d'ajuster l'amplitude du signal transformé. On utilise la TFCT pour passer dans le domaine fréquentiel ce qui va nous permettre d'étudier et de faire des modifications de manière plus simple sur les signaux.

La fonction PVoc appelle alors la fonction TFCT qui prend en paramètre le signal (x), le 'Nfft', 'Nwind' et 'Nov' qui ont été décrits précédemment.
On applique alors à la sortie le facteur d'échelle (ici : scf=1).

Une fois dans la fonction, on a besoin de s'assurer que les données de x (le signal) soient bien sous forme d'une matrice d'une seule ligne (un vecteur) puis on récupère le nombre d'échantillons de ce vecteur qu'on appellera N ainsi qu'on initialise une valeur nommée 'c' qui permettra de suivre l'indice de la colonne résultante D. Maintenant nous sommes prêts à appliquer la fenêtre de Hanning, cette fenêtre est également appelée fenêtre de cosinus levé. Cette fenêtre de pondération donc la formule mathématique est la suivante :

$$w(n) = 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right)$$

Fig 3 : Formule fenêtre de cosinus levé

Cette fenêtre est utilisée pour la réduction de lobes secondaires. En effet, quand on effectue une TF sur un signal fini des lobes secondaires peuvent apparaître dans le spectre et peuvent altérer la précision de l'analyse fréquentielle. De plus, lorsqu'on divise un signal d'origine en trames pour effectuer une TFCT il peut y avoir des artéfacts créés par le découpage du signal original. C'est pour cela qu'ajouter une fenêtre de pondération est importante. Les lobes secondaires sont marqués par des bandes de fréquences qui apparaissent dans le spectre lorsqu'il est transformé dans le domaine fréquentiel. Après avoir créé la fenêtre de pondération nous pouvons calculer la TFCT et renvoyer une matrice résultante. On applique la TFCT au signal x en utilisant la fenêtre de Hanning pour pondérer chaque segment du signal. Le résultat est stocké dans la matrice D, où chaque colonne représente le spectre fréquentiel d'une portion du signal x (chaque colonne représente les échantillons de la TF d'une trame), tous les éléments sont complexes car ils sont issus de la TF donc ont un module et une phase.

Après avoir appelé la fonction TFCT nous allons appeler la fonction TFCT_interp. La fonction prend en paramètre le "nouveau" signal X qui a été renvoyé à l'étape précédente, une nouvelle base de temps qui est notée 'Nt' qui espace les indices de la colonne de X d'un pas 'rapp' la valeur vue au début. On utilise le même nombre d'échantillons de chevauchement 'Nov' que précédemment. Cette fonction réalise l'interpolation temporelle.

Cela permet d'obtenir ainsi un signal le plus fidèle possible à l'audio original lorsqu'on reconstruit le signal à la fin en utilisant la TFCT inverse. Il faut qu'en chaque fréquence la phase du signal modifié soit approximativement la même que celle du signal original. En effet, l'information qui est présente entre 2 trames est également importante c'est pourquoi on doit faire une interpolation du spectre d'amplitude et un traitement particulier de la phase. On effectue une interpolation linéaire du spectrogramme X aux indices temporels spécifiés par le vecteur « t » qui a été pris en paramètre de la fonction TFCT_Interp. L'interpolation est effectuée en utilisant des coefficients linéaires calculés en fonction des parties fractionnaires des indices temporels. Les phases interpolées sont ajustées pour maintenir la continuité du signal et les résultats sont stockés dans la matrice 'y'. Si ce n'était pas le cas, il y aurait du bruit qui serait dû aux hautes fréquences induites lors des ruptures de phase.

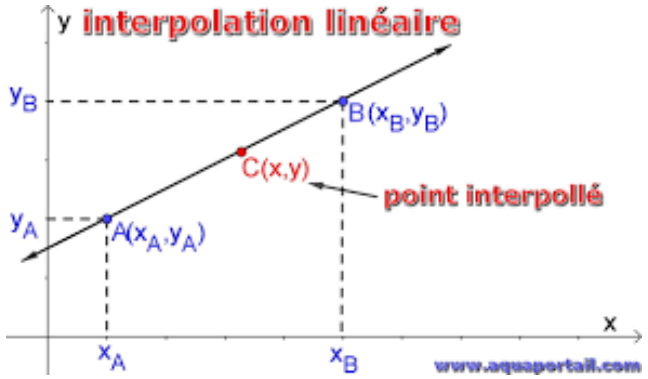

Fig 4 : Schéma représentant ce qu'est l'interpolation linéaire Aquaportail)

Puis pour finir, nous effectuons la Transformée de Fourier inverse, ce qui nous permet de retrouver un signal dans le domaine temporel et de pouvoir l'écouter. Une fenêtre de pondération est utilisée pour compenser le chevauchement des fenêtres.

Pour étudier l'impact du PVOC sur le signal initial, nous allons comparer le signal initial et le signal reconstruit sans faire aucune transformation supplémentaire

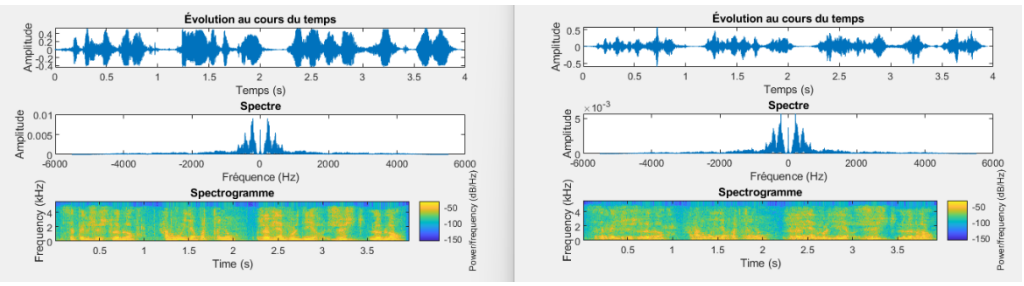

Fig 5 : Comparaison signal original et signal Pvoc (sans accélération ou ralentissement)

La fonction Pvoc modifie le signal original même en gardant la durée. L'utilisation de la TFCT, de la fenêtre de Hanning, de l'interpolation et de la TFCT inverse ne permet pas de retrouver le signal initial tel quel.
Une partie de l'amplitude est plus faible. Cependant, le spectre global et le spectrogramme restent similaires.

En résumé, nous avons eu un signal x(t) que nous avons découpé en N trames. Chaque trame appelée i est dite “centrée” à l’instant txi. Chaque trame va subir une Transformée de Fourier, qui va être rangée dans la matrice X, chaque colonne va représenter un échantillon; on aura donc N colonne.
Ensuite, pour l’interpolation on utilisera donc X comme paramètre de la fonction d’interpolation ainsi que tyo . POur obtenir le signal y(t) on va transformer la matrice X en matrice Y de M colonnes (trames) où chaque colonne est la TF d’une trame de y(t) centrée en tyi. tyi est le vecteur temps sur lequel on interpole. Pour obtenir les valeurs de la colonne Y, nous allons effectuer des calculs basés sur deux colonnes spécifiques de X. Le choix de ces colonnes de X dépendra de la position relative de la colonne Y parmi les colonnes de X. Nous répéterons ce processus pour chaque colonne de Y, calculant ainsi les valeurs correspondantes à partir des colonnes appropriées de X.
Exemple :

Calcul de y1 en ty1 :
y1 = ½ x1 + ½ x2 soit ty1 ϵ [$tx1$, $tx2$]
ty1 au centre de tx1 et tx2 ⇒ les facteurs ½

Calcul de y2 en ty2 :
$y2$ = α. $x1$ + β. $x2$ soit ty2 ϵ [$tx1$, $tx2$]

α : en fonction de la position de ty2 par rapport à tx1
β : en fonction de la position de ty2 par rapport à tx2

On fait cela pour toutes les colonnes (voire le schéma à la page suivante)

Schématisation de l'algorithme :

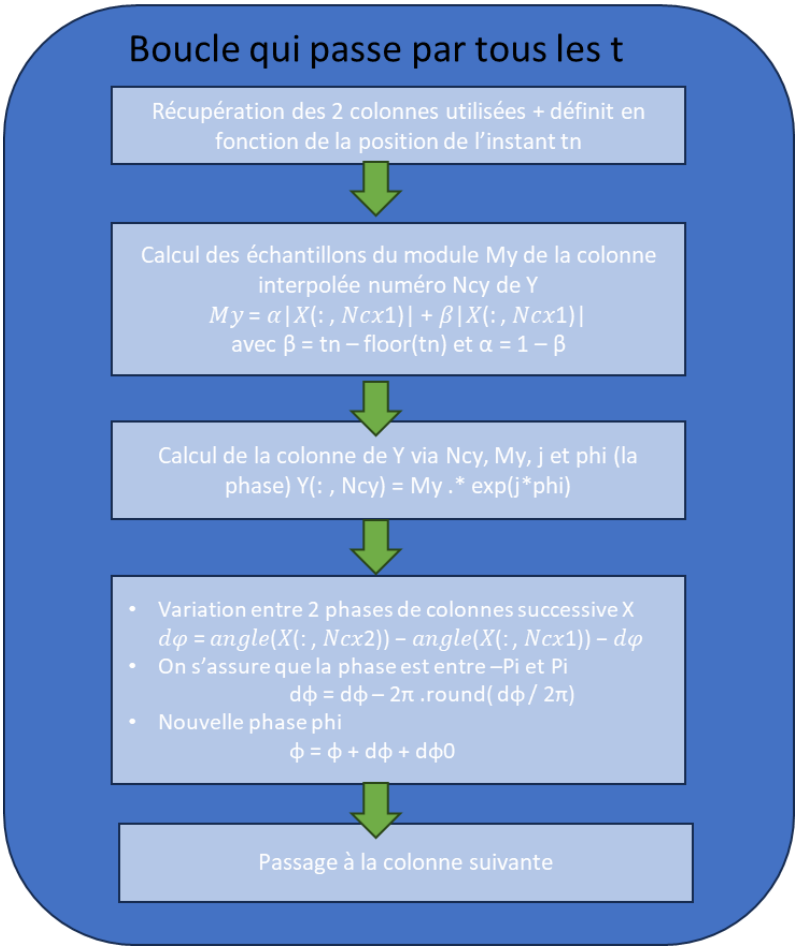

Fig 6 : Représentation schématique de ce que fait l'algorithme d'interpolation

Puis on appelle la fonction TFCT_inv qui nous ramène dans le domaine temporel.

2.3 Ralentissement et accélération :

On appel tout simplement PVoc en lui passant en paramètre les informations présentées précédemment :

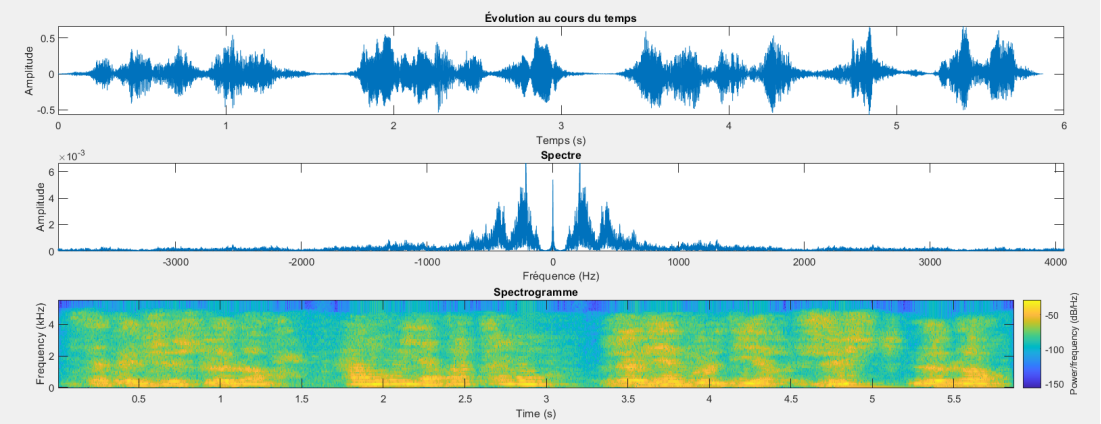

Fig 7 : Représentation du signal Dinner.wav après ralentissement dans les domaines temporels, fréquentiels et le Spectrogramme (rapp = ⅔)

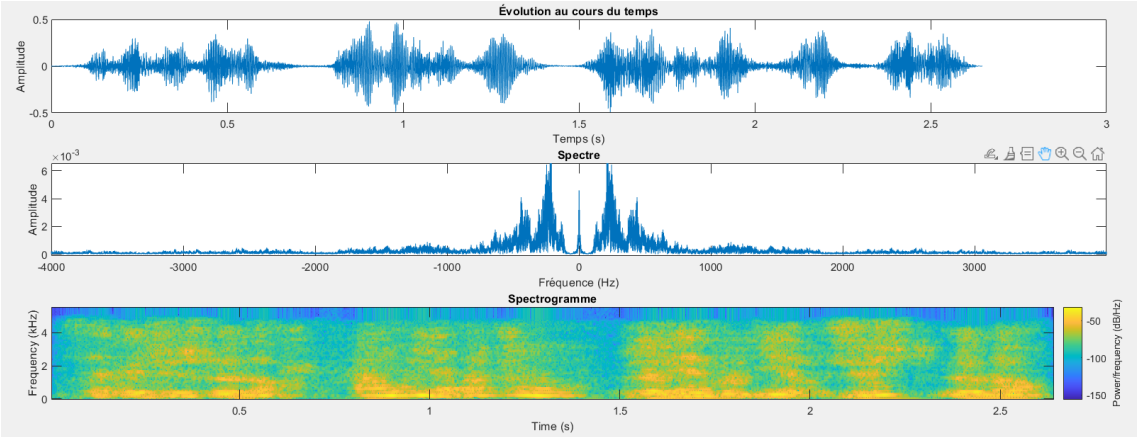

Fig 8 : Représentation du signal Dinner.wav après accélération dans les domaines temporels, fréquentiels et le spectrogramme (rapp = 3/2)

Domaine temporel :

On observe que les sons ont des durées différentes, ce qui est attendu puisqu'ils ont été ralentis ou accélérés. Bien que les amplitudes des sons soient presque similaires, certaines sont "écrasées". Cela se produit soit parce que le temps est plus court, nécessitant ainsi le passage de plus de fréquences dans une période plus courte, soit parce que le temps est plus long, obligeant à répartir le même "nombre" de fréquences sur une période plus étendue.

Spectre :

Nous remarquons que les deux spectres et le spectre original se ressemblent grandement et qu'ils semblent avoir une amplitude identique. De plus, la gamme de fréquences et la fréquence fondamentale n'ont pas été modifiées. On a donc bel et bien modifié la voix sans

modifier le pitch car le pitch est caractérisé par une certaine fréquence fondamentale (aiguë ou grave).

Spectrogramme :

Il est observé que les pics d'énergie se dispersent différemment en fonction du taux d'extraction. Dans le spectre du signal ralenti, les pics d'énergie sont plus étalés que dans le signal d'origine. En revanche, pour l'accélération, les pics d'énergie sont plus resserrés, créant ainsi une compression dans le domaine fréquentiel. Ces observations reflètent les changements dans la densité spectrale de puissance des signaux, résultant de la modification de leur vitesse, et sont mises en évidence dans les différents spectrogrammes générés par le code, à savoir le signal original, le signal ralenti et le signal accéléré.

2.4 Augmentation et diminution du pitch sans modification de la vitesse

La modification du pitch permet de garder la même durée du signal et de rendre la voix plus aigüe ou plus grave.

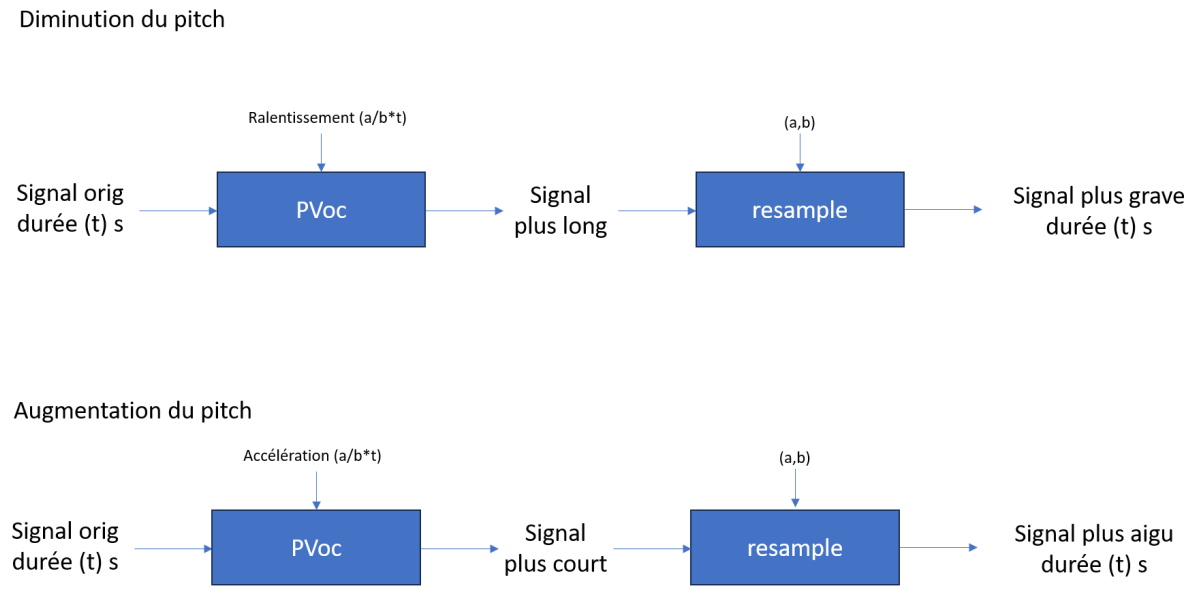

Fig 9 : Processus pour rendre les voix plus aiguës ou plus graves

Pour augmenter ou diminuer le pitch, on appelle PVoc dont nous avons décrit l'utilisation précédemment et nous allons ralentir le signal initial (pour augmenter le pitch) ou ralentir le signal de départ (pour diminuer le pitch). Puis avec le résultat obtenu on ré-échantillonne avec la fonction de matlab ‘resample’ afin de faire en sorte que ce nouveau signal fasse la même durée que le signal d’origine.

La fonction resample y = resample(x,p,q) rééchantillonne la séquence d'entrée, x, à p/q fois la fréquence d'échantillonnage d'origine. Elle permet donc de changer la fréquence d’échantillonnage d’un signal en enlevant ou en rajoutant des échantillons. Le rééchantillonnage applique un filtre à réponse impulsionnelle finie (FIR) à x et compense le retard introduit par le filtre. La fonction fonctionne le long de la première dimension du tableau dont la taille est supérieure à 1 ce qui est notre cas.

Ainsi, on modifie la vitesse du pitch en augmentant ou en diminuant les fréquences puis après on modifie la durée du signal pour qu’elle corresponde avec celle du signal d’origine.

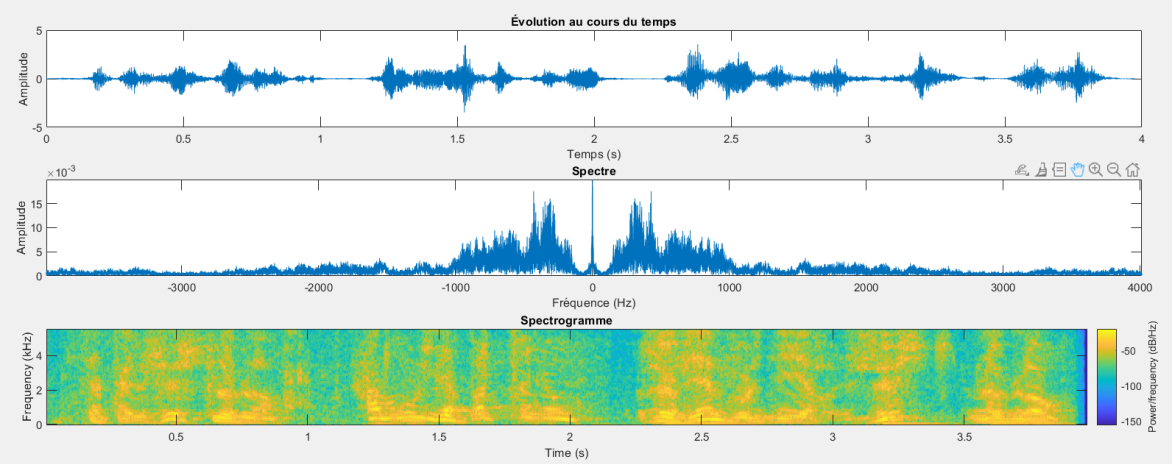

Fig 10 : Représentation du signal Dinner.wav après Augmentation du pitch dans les domaines temporels, fréquentiels et le Spectrogramme (a=2, b=3)

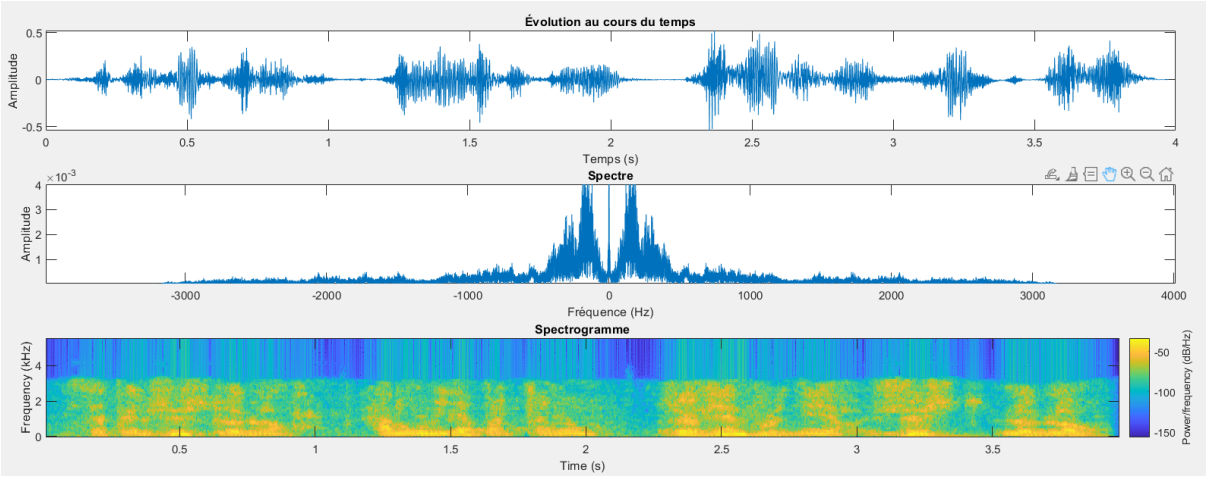

Fig 11 : Représentation du signal Dinner.wav après Diminution du pitch dans les domaines temporels, fréquentiels et le Spectrogramme (a=3, b=2)

Domaine temporel :

On remarque qu'il y a eu un changement d'amplitude lors de l'augmentation ou la diminution du pitch. On constate qu'une augmentation de l'amplitude conduit à une augmentation du signal et qu'une diminution de l'amplitude entraîne une diminution du signal.

Domaine fréquentiel :

On observe les spectres attentivement et on peut remarquer que l'augmentation du pitch atténue les fréquences les plus faibles et amplifie les fréquences les plus élevées correspondantes à la voix. C'est pourquoi lors de l'augmentation du pitch on perçoit une voix plus aiguë. A l'inverse, lors d'une diminution les fréquences les plus grandes correspondantes à une voix sont atténuées mais les fréquences les plus faibles sont augmentées. Par conséquent on perçoit des voix beaucoup plus graves.

Spectrogramme :

Comme dit précédemment lors de l'analyse du spectre, l'augmentation du pitch mène à une voix plus aiguë. On peut le voir en regardant les pics d'énergie qui sont représentés par la partie jaune dans le spectrogramme, qui s'étendent jusqu'à 5 KHz. Au contraire, la diminution du pitch mène à un son plus grave qui se traduit clairement par des pics d'énergie qui ne dépassent pas les 3 KHz.

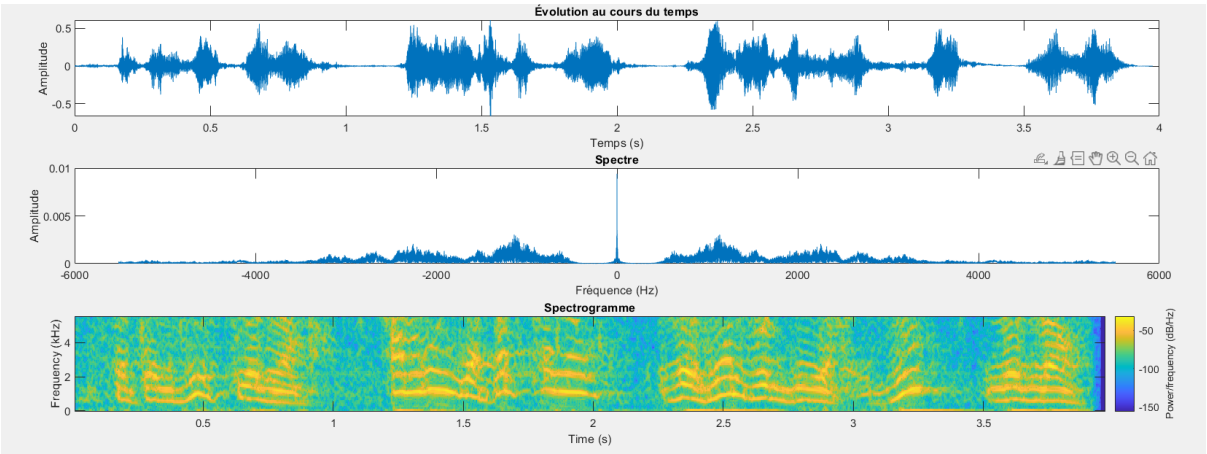

Fig 12 : Représentation du signal Dinner.wav après Augmentation du pitch dans les domaines temporels, fréquentiels et le Spectrogram (a=1, b=5)

Si on choisit on choisit un rapport de a/b trop petit alors la voix sera extrêmement aiguë et nous ne pouvons plus distinguer ce qui est dit lors de l'audio.

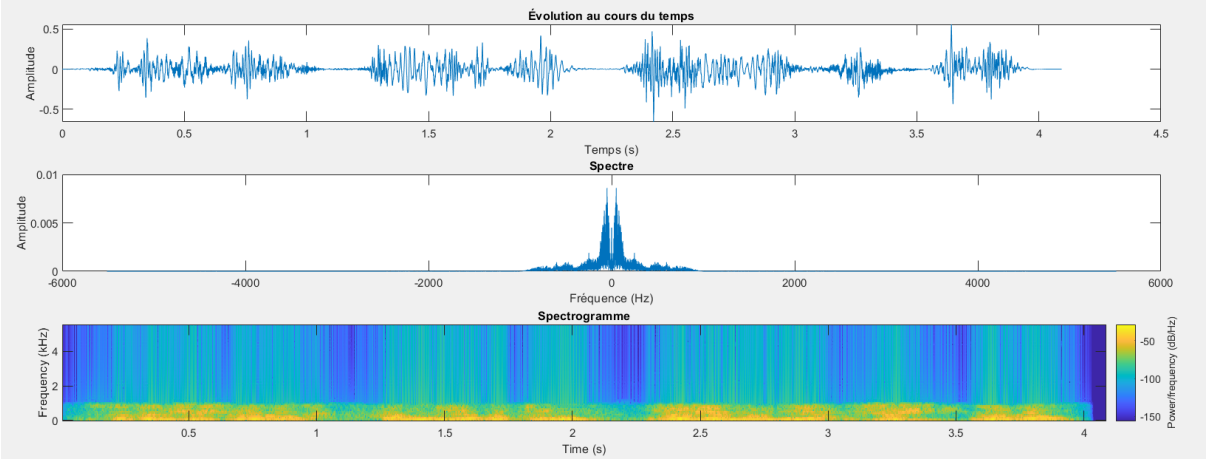

Fig 13 : Représentation du signal Dinner.wav après Diminution du pitch dans les domaines temporels, fréquentiels et le Spectrogram (a=5, b=1)

A l'inverse, si on choisit un rapport de a/b trop grand alors la voix sera extrêmement grave et nous ne pouvons plus distinguer ce qui est dit lors de l'audio.

3 Partie B : Robotisation de la voix

La robotisation d'une voix permet d'obtenir un son plus synthétique, permet de faire parler des robots comme dans Star Wars. Cette transformation se fait dans le domaine temporel contrairement aux autres modifications faites précédemment. Pour ce faire nous allons créer la fonction Rob qui va modifier le signal audio dans le domaine temporel en le modulant avec une exponentielle complexe à une fréquence notée 'Fc'.

On commence donc par définir un vecteur temporel noté 't', le vecteur temporel permet de représenter chaque échantillon dans le temps. Quand un signal analogique est converti en signal numérique, il est divisé en petits morceaux appelés échantillons. Chaque échantillon est enregistré à des intervalles réguliers de temps. La fréquence d'échantillonnage Fs est le nombre d'échantillons par unité de temps. En divisant chaque indice d'échantillon par la fréquence d'échantillonnage nous obtenons une échelle de temps dans laquelle chaque élément du vecteur temporel représente un instant dans le temps, exprimé en secondes.

Ensuite, on module le signal avec une exponentielle complexe yrob = y(t)* $e^{-j*2*Pi*Fc*t}$
Pour finir, on récupère la partie réelle de ce signal car c'est lui qui constitue le signal robotisé. Le paramètre Fc doit être bien choisi car il détermine la fréquence à laquelle la voix va être robotisée. Si on choisit des valeurs de Fc trop grandes, cela peut créer un effet sonore très intense et rendre le signal inaudible.

Ainsi, la fonction Rob transforme le signal audio en le modulant avec une exponentielle complexe, puis en extrayant sa partie réelle. Cela crée un effet sonore de robotisation. Les valeurs précises de Fc déterminent l'intensité de cet effet robotique et peuvent être ajustées pour obtenir le résultat sonore désiré.

La transformée de Fourier du nouveau signal sera donc : TF(yrob)=TF(f+Fc).
Le spectre a la même forme que le signal mais avec un décalage de fréquence Fc.

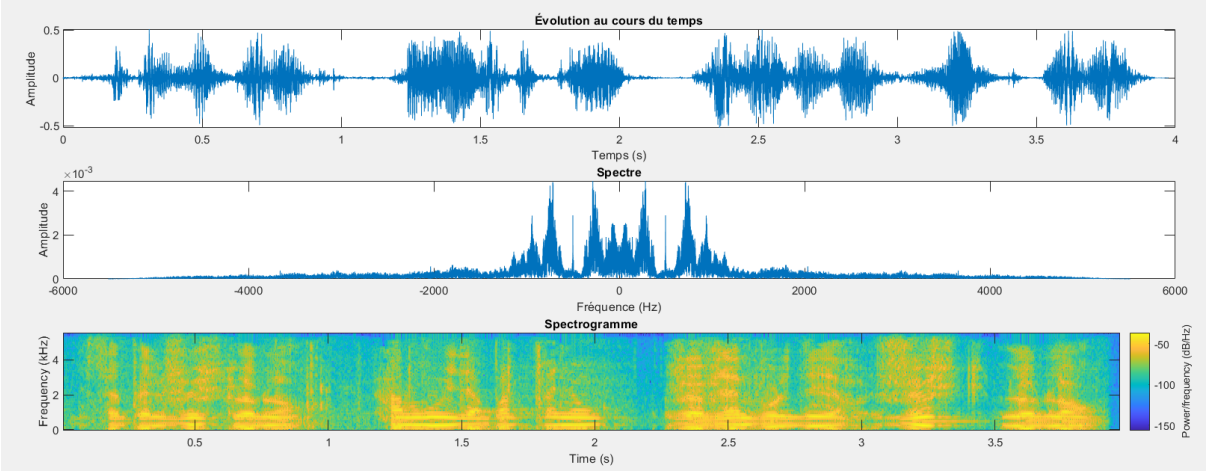

Fig 14 : Représentation du signal Dinner.wav après Robotisation du pitch dans les domaines temporels, fréquentiels et le Spectrogramme (Fc=500)

On observe que la gamme de fréquence a été décalée de la valeur Fc, plus Fc est grand plus le décalage est important. Ceci s'explique par le fait que la robotisation s'effectue à l'aide d'une exponentielle complexe. On peut constater sur les spectrogrammes que les pics de fréquence n'ont pas étés modifié, ils ont simplement été décalés autour de la fréquence porteuse. On a simplement modifié la fréquence fondamentale du son.
Il se trouve que l'oreille humaine comprend toujours le son tant que les fréquences sont inférieures à 2000Hz.

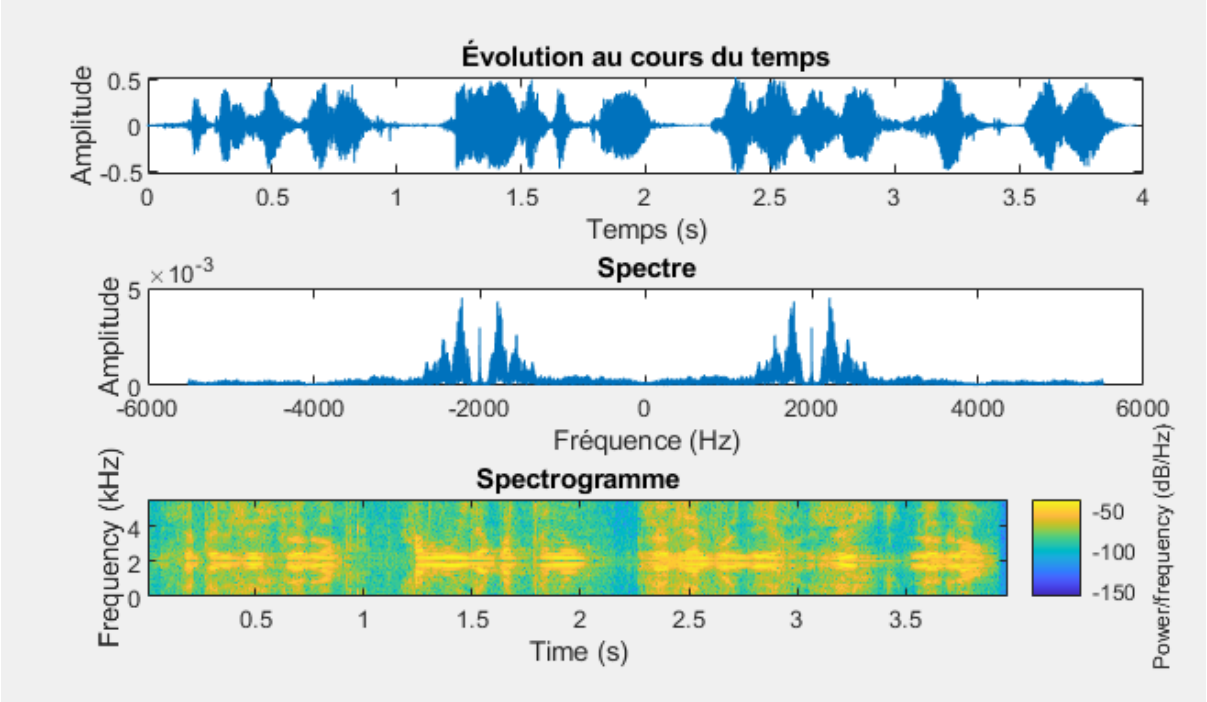

Fig 15 : Représentation du signal Dinner.wav après Robotisation du pitch dans les domaines temporels, fréquentiels et le Spectrogram (Fc=2000)

Si l'on choisit une modulation de 2000Hz, l'oreille humaine ne capte plus les fréquences au-delà de cette valeur et après écoute, comme on pouvait s'en douter avec le choix de Fc = 2000 Hz, l'audio devient inaudible.

4 Conclusion :

Durant ce projet nous avons mis en place des techniques et des méthodes d'approches vue en cours. Nous avons également dû nous renseigner et nous former sur ce qu'est la robotisation et l'interpolation d'un signal. Nous avons également pu nous amuser à tester les différents "filtres" sur notre propre voix ou sur des audios des films/séries. Cela nous a permis de mieux comprendre le principe des effets spéciaux sur la voix qui sont utilisés dans les films.
Nous avons pu constater que modifier la vitesse d'un signal influe sur le caractère extensible d'un signal, plus le son est rapide plus les ondulations seront concentrées, et inversement. En utilisant le rapport a/b, nous avons pu amplifier/réduire les amplitudes des harmoniques afin d'impacter la fréquence globale et donc la vitesse.
Pour la modification du pitch, nous avons utilisé la même méthode que pour la modification de la vitesse, sauf qu'à la fin il a fallu rétablir la vitesse initiale. Cette fois-ci, la fréquence de la fondamentale s'est vue modifiée et cela à eu un impact sur les harmoniques. Dans le cadre d'une augmentation, les hautes fréquences sont amplifiées et on obtient un son plus aigu, et inversement dans le cas d'une diminution.
Pour la robotisation, nous avons constaté l'impact de la fréquence de coupure Fc sur la robotisation. En effet, plus Fc est élevé plus le son initial était robotisé.
Le passage du domaine temporel au domaine fréquentiel permet de faire de nombreux calculs qui ne sont pas possibles dans le domaine temporel uniquement. On peut ainsi modifier l'amplitude de chacune des fréquences en plus des modifications qu'on a fait précédemment.

5 Bibliographie :

Transformée de fourier : https://moodle.luniversitenumerique.fr/course/view.php?id=14

Transformée de fourier matlab :
https://www.electronique-mixte.fr/matlab-transformee-de-fourier/

Transformée de fourier inverse matlab :
https://fr.mathworks.com/help/matlab/ref/ifft.html?s_tid=mwa_osa_a

Fenêtre de Hanning :

- https://fr.mathworks.com/help/signal/ref/hann.html
- http://sti-monge.fr/maintenance/maintenance_a1z2e3/test_maintenance/files/05_fr_outils_traitement.pdf

Interpolation linéaire : https://www.aquaportail.com/dictionnaire/definition/2026/interpolation

Lobes secondaires : http://users.polytech.unice.fr/~leroux/courssignal/node74.html

MathWorks (Matlab) functions: https://fr.mathworks.com/help/signal/ref/resample.html