Introduction à la reconnaissance de la parole

Gabriel Dauphin

September 11, 2024

Contents

Introduction

Ce cours présente un ensemble de techniques qui sont utilisées en travaux pratiques pour distinguer les sons correspondant aux mots `un`, `deux` et `trois`. Ces techniques utilisent la notion de description temps-fréquence qui est implicite dans la décomposition d'un son en trames, explicite dans la notion de `spectrogramme` et qui est utilisée dans les différents descripteurs présentés dans le chapitre 3. Le chapitre 1 a pour objectif de décrire la façon des sons sont représentées, cela amène à décrire brièvement les organes humains de la perception et le prétraitement qui est appliqué aux données associées aux sons. Le chapitre 2 a pour objectif de présenter la classification, à savoir son objectif, la façon dont on l'évalue et la manière dont elle utilise des valeurs calculées pour chaque son que l'on appelle ici *descripteur*. Le chapitre 3 présente différentes techniques permettant de calculer les valeurs de descripteurs à partir des données associés aux sons. Cela amène à décrire brièvement les organes humains de la génération d'un son car cela explique l'intérêt porté à l'utilisation d'une description fréquentielle de portions de signaux temporels.

En tant que document de cours, il a pour objectif de préciser les notions à connaître. Les notions à connaître le plus précisément sont la compréhension et la connaissance des *formules encadrées*, la compréhension des *graphiques et schémas* présentés et la connaissance et la signification des *termes techniques*.

Chapter 1

Représentation d'un son

Pour décrire et discuter de la façon dont les sons sont représentées, la section 1.1 discute de ce qu'est physiquement un son et donne une brève description des organes humains utilisées pour la perception d'un son. La section 1.2 rappelle comment le traitement numérique du signal permet de représenter un son. La section 1.3 montre une façon un peu différente de représenter le son au moyens de *trames*, c'est-à-dire de petites portions du signal sonore. La section 1.4 montre une façon simplifiée de détecter les phases de silence et de paroles, il se trouve que ceci était un des objectifs de ce cours. Cette détection est utilisée pour réaliser un prétraitement permettant d'améliorer la performance de la classification. La section 1.5 donne quelques indications sur la façon dont les techniques proposées sont implémentées dans les travaux pratiques qui illustre ce cours.

1.1 Qu'est-ce qu'un son ?

La section 1.1.1 présente une description du son en tant qu'onde sonore se propageant dans l'air. La section 1.1.2 présente une très brève description du fonctionnement de l'oreille en tant qu'organe de perception des sons. La section 1.1.3 présente des courbes d'audibilité, c'est-à-dire les seuils hauts et bas de perception du son.

1.1.1 Modélisation physique

Un son est généralement décrit comme plus ou moins aigu ou plus ou moins fort. Nous allons voir que la première notion correspond à une fréquence et la deuxième à un niveau sonore.

Le son est une variation de la pression de l'air. [1] Cette variation se propage sous la forme d'une onde. Une modélisation sphérique de cette onde est décrite par

$$P(r,t) = P_0 + \frac{1}{r}\left[g_1(t-\frac{r}{c}) + g_2(t+\frac{r}{c})\right] \tag{1.1}$$

- P_0 est la pression atmosphérique moyenne. L'unité dans laquelle elle s'exprime est le Pa (Pascal). La pression de l'air est environ 10^5Pa.
- r est la distance en mètres (m) entre le point où on mesure la pression et la source de ce signal sonore.
- t est l'instant en secondes (s) pour lequel on mesure la pression $P(r,t)$.
- c est la vitesse (m.s^{-1}) de propagation du son à travers l'air. C'est environ de 340m.s^{-1}.
- g_1 et g_2 peuvent a priori être des fonctions quelconques. Pour modéliser la propagation d'un signal sonore, on considère $g_2 = 0$ et pour un son stationnaire par exemple dans une caisse de résonnance, g_1 et g_2 sont similaires.

Pour modéliser une note de musique qui dure un temps infini, on utilise

$$\boxed{g_1(\tau) = \cos(2\pi f_0 \tau)} \tag{1.2}$$

[1] Un son peut aussi se propager à travers d'autres matériaux et dans ce cas c'est le matériau traversé qui se met à vibrer, en revanche le son ne peut pas se propager à travers le vide.

Cette note est caractérisée par sa fréquence f_0 en Hz. En effet, avec cette modélisation on observe que $P(r, t + \frac{1}{f_0}) = P(r, t)$, c'est-à-dire qu'en chaque point la pression évolue dans le temps avec une périodicité de $\frac{1}{f_0}$.

$$P(r, t + \frac{1}{f_0}) = P_0 + \frac{1}{r}\cos(2\pi f_0(t + \frac{1}{f_0} - \frac{r}{c})) = P_0 + \frac{1}{r}\cos(2\pi f_0(t - \frac{r}{c})) = P(r, t)$$

Et f_0 est reliée à une longueur d'onde notée λ_0 mesurée en mètres.

$$\boxed{\lambda = \frac{c}{f_0}} \tag{1.3}$$

Ceci traduit une évolution quasi-périodique du signal dans l'espace $P(r + \lambda, t) \approx P(r, t)$.

$$P(r + \lambda, t) = P_0 + \frac{1}{r + \lambda}\cos(2\pi f_0(t - \frac{r}{c} - \frac{\lambda}{c})) = P_0 + \frac{1}{r + \lambda}\cos(2\pi f_0(t - \frac{r}{c})) \approx P(r, t)$$

À l'oreille, ce son apparaît comme une note de musique. Il s'agit d'un La lorsque $f_0 = 440$Hz. Cette fréquence est donc liée au caractère plus ou moins aigu du son.

L'intensité sonore d'un son correspond au produit de la surpression par la vitesse. Dans le cadre de ce cours, on considère qu'il s'agit du carré de la partie variable de la pression.

$$I = (P(r, t) - P_0)^2$$

Cette intensité sonore est en Watt par mètres carrés (W.m^{-2}). Dans le cadre de ce cours, on attache de l'importance au fait que cette intensité sonore est proportionnelle à $\frac{1}{r^2}$.

$$\boxed{I = \frac{I_e}{r^2}} \tag{1.4}$$

I_e étant une quantité inconnue à estimer. Aussi lorsqu'on s'éloigne de la source en doublant la distance, l'intensité sonore est divisée par 4.

On appelle niveau sonore, le fait d'utiliser une échelle en décibel pour mesurer l'intensité sonore.

$$\boxed{L = 10\log_{10}\left(\frac{I}{I_0}\right)} \tag{1.5}$$

où I_0 est une intensité sonore de référence. On peut remarquer que dans l'équation (1.5), l'échelle en décibel n'est pas calculé avec un facteur multiplicatif de 20 mais de 10, parce que l'intensité sonore est déjà un terme carré. Et en effet lorsqu'on s'éloigne d'une source sonore en multipliant par deux la distance, le niveau sonore est diminué de $10\log_{10}(4) = 6$dB. [2]

1.1.2 Fonctionnement schématique de l'oreille

La figure 1.1 (extraite d'une publication) représente un schéma détaillé de l'oreille. Schématiquement, le son s'introduit dans le conduit auditif, fait vibrer le tympan qui actionne le marteau, l'enclume et l'étrier qui à son tour fait vibrer le liquide présent dans la cochlée. Cette cochlée est tapissée de cellules ciliées qui transmettent les vibrations vers des nerfs auditifs. Les cellules au début de la cochlée sont sensibles aux sons aigus tandis que celles en fin de cochlée sont sensibles aux sons graves.

1.1.3 Courbes d'audibilité humaine

Dans la figure 1.2 (extraite d'une publication), l'échelle des ordonnées est le niveau sonore d'un son écouté. L'échelle des abscisses est une échelle de fréquences (Hz) mais avec une distribution logarithmique. À gauche ce sont les infrasons qui correspondent aux sons qu'on ne peut entendre parce que leur fréquence est trop basse. À droite ce sont les ultrasons qu'on ne peut pas non plus entendre parce que leur fréquence est trop haute. Entre ces deux fréquences, il y a deux courbes, si le niveau sonore d'un son est supérieur à la limite donnée par la courbe inférieure alors on peut l'entendre et s'il est supérieur à la limite donnée par la courbe supérieure alors ce son peut abîmer

[2] C'est en effet la même diminution que pour le rapport signal sur bruit, si on divisait par deux le bruit sans modifier le signal source.

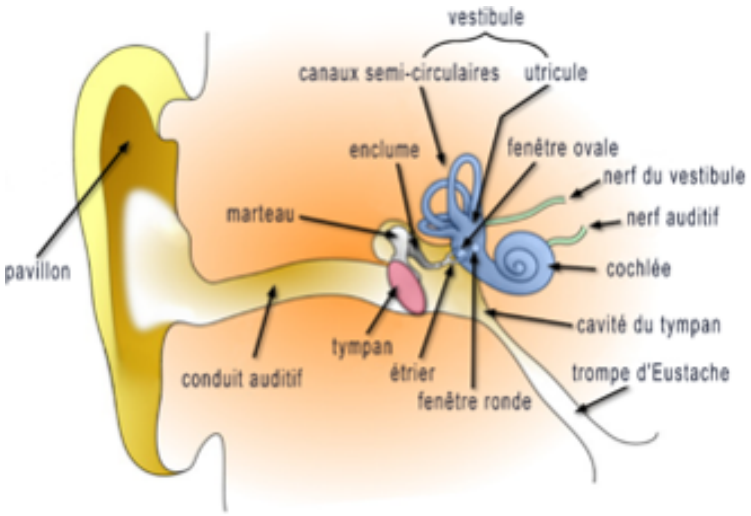

Figure 1.1: Schéma détaillé de l'oreille.

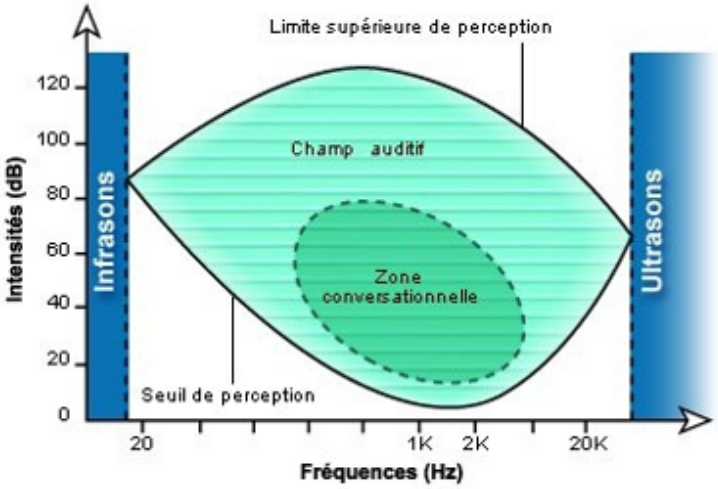

Figure 1.2: Courbe d'audibilité humaine.

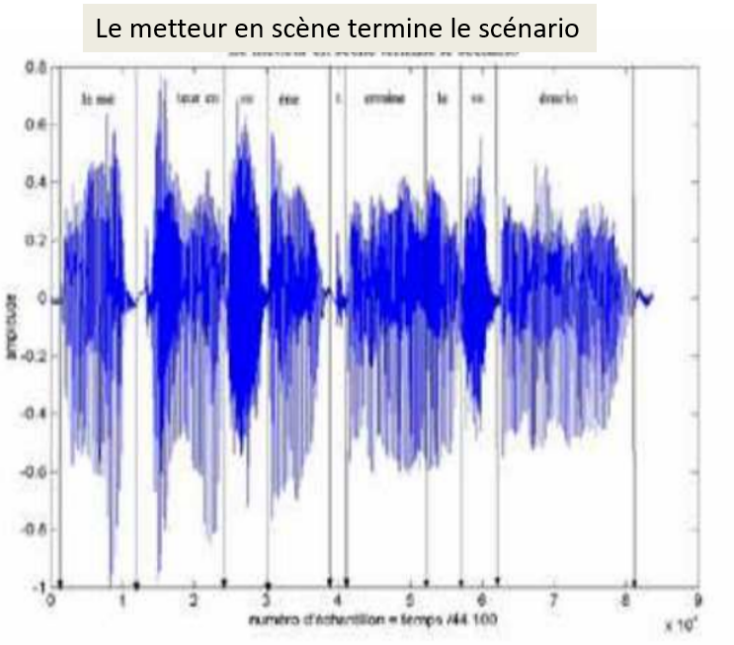

Figure 1.3: Représentation temporelle du signal sonore.

l'oreille. Ces courbes sont obtenues en expérimentant sur un grand nombre de patients, elles indiquent un sensibilité moyenne, mais il y a en réalité des différences importantes d'un individu à un autre. Ces courbes nous amènent à affirmer qu'une assez grande partie de l'information sonore est obtenue en ne considérant que les fréquences inférieures à 4kHz et qu'a priori la totalité de l'information sonore est obtenue en ne considérant que les fréquences inférieures à 20kHz.

1.2 Représentation temporelle d'un son

En tant qu'onde sonore, un son est un signal temps continu à valeurs continus. Mais lorsqu'il est converti en signal électrique avec un capteur acoustique appelé microphone puis converti en signal numérique avec échantillonnage et quantification, le signal devient temps discret. Du fait du critère de Shannon-Nyquist, la fréquence d'échantillonnage f_e est entre 40kHz et 50kHz (en général 44.1kHz mais, elle peut parfois être de 8kHz. Le signal est souvent quantifié sur 8 bits, mais cela peut être 10 ou 12 bits. Le signal échantillonné est noté x_n, chaque indique n correspondant à un instant $t_n = nT_e$ où $T_e = \frac{1}{f_e}$. La figure 1.3 (extraite d'une publication), indique l'évolution au cours du temps de l'intensité sonore. En ordonnée, il n'y a pas d'unité, c'est a priori une fraction de la valeur maximale. Mais le carré du signal est proportionnel à l'intensité sonore. Le signal représenté est celui obtenu en enregistrant la phrase *Le metteur en scène termine le scénario.*

On appelle *durée* du signal, le temps qui s'écoule entre le début du signal sonore et la fin. Cette durée est égale à NT_e où ici N est le nombre d'échantillons [3] représenté sur la figure 1.3. Étant donné la valeur élevée de f_e, on a en fait un très grand nombre d'échantillons, bien plus qu'il n'est possible de distinguer à l'oeil en observant la figure. L'allure générale permet aussi de distinguer les moments où il y a du silence, des moments où il y a un signal sonore important : le silence correspond à des valeurs très faibles en valeur absolues et le signal sonore important correspond à des valeurs importantes du signal. Comme chaque intervalle de temps dans le graphe correspond aussi à un intervalle de temps durant lequel un son est prononcé, il est possible de faire correspondre chaque syllabe à un morceau du signal. On peut ainsi visualiser pour certaines sons des particularités, le premier t et les deux s correspondent à des parties très foncées du graphe.

Bref, cette représentation temporelle du signal suggère l'étude séparée de chaque morceau de signal correspondant à un son ou à une syllabe. Mais comme on ne connaît pas a priori quand commence et quand finit le son ou la syllabe, la section 1.3 montre que l'on découpe le signal en une succession de trames.

Les notations utilisées pour décrire l'ensemble du signal sont les suivantes.

[3] Étant donné la valeur de f_e, il est tout à fait satisfaisant d'approcher cette durée à un multiple de T_e.

- T durée du signal en secondes.
- N nombre total d'échantillons.
- $t_n = nT_e$ échelle de temps du signal pour $n \in \{0 \dots N-1\}$.

On a la relation entre durée du signal et période d'échantillonnage.

$$\boxed{T = NT_e} \tag{1.6}$$

On a aussi l'échantillonnage du signal.

$$\boxed{x_n = x(t_n)} \tag{1.7}$$

Pour illustrer des notions traitement du signal, on s'interroge ici sur la façon de rendre un son plus grave au moyen d'un traitement numérique du vecteur x_n.

- En diminuant f_e, le son est plus grave. Mais sa durée est aussi diminuée d'autant puisqu'elle est égale à $\frac{N}{f_e}$.
- Si on effectue un sur-échantillonnage tout en conservant la fréquence d'échantillonnage alors le son devient plus grave mais sa durée est augmentée d'autant.
- Si on effectue une modulation, c'est-à-dire qu'on multiplie le graphe par une sinusoïde d'une certaine fréquence alors on déplace effectivement le son à la fois en le rendant plus aigu et plus grave. Si ce son a un spectre très étroit alors il est possible de supprimer la composante aigüe ou la composante grave. Et la durée est alors inchangée.

1.3 Représentation du signal sonore en une succession de trames

1.3.1 Découpage en trames sans chevauchement

La figure 1.4 donne quatre représentation d'un même signal défini par

$$x_n = \sin(20\pi n T_e)$$

C'est le signal représenté en haut à gauche de cette figure où sur une seconde, il y 10 périodes de la sinusoïde qui sont répétées et où donc, la fréquence de la sinusoïde est de 10Hz et la période $T = 0.1$s. L'échelle des ordonnées est une fraction de l'amplitude maximale. L'échelle des abscisses est la seconde.

Dans cet exemple, on découpe le signal d'une seconde en 10 trames chacune dure ici 0.1s. La figure en haut à droite représente chaque trame avec une couleur différente : bleu pour la première trame, vert pour la deuxième, rouge pour la troisième, cyan pour la quatrième, violet pour la cinquième, etc...

Le graphe en bas à gauche est une autre représentation de ces trames. L'axe pointant vers la droite est le temps écoulé depuis le début de chaque trame. L'axe pointant vers la gauche et commençant à zéro au milieu indique le numéro de la trame considérée. L'axe vers le haut indique la valeur du signal pour la trame considérée à l'instant de la trame considérée. La valeur zéro est ici à mi-hauteur. Ici la couleur ne distingue plus les trames entre elles, elle dépend de la valeur prise par le signal. Dans cet exemple le signal est identique sur chaque trame, c'est pour cela qu'on retrouve exactement le même signal sur chaque courbe.

La courbe en noir du graphe en bas à droite est la même que le graphe en haut à gauche, il s'agit du signal x_n représenté en fonction du temps. Sur ce graphe, il y a en plus des traits horizontaux de couleurs et de hauteurs différentes qui permettent d'identifier chaque trame. Ce sont ces traits qui permettent de lire les instants associés au début et à la fin de chaque trame en trouvant l'abscisse du début et de la fin du trait horizontal.

1.3.2 Notations pour définir les trames

On utilise ici les notations suivantes. Ces notations ne sont pas universelles, il est donc nécessaire de les redéfinir quand on les utilise.

- K désigne le nombre de trames.

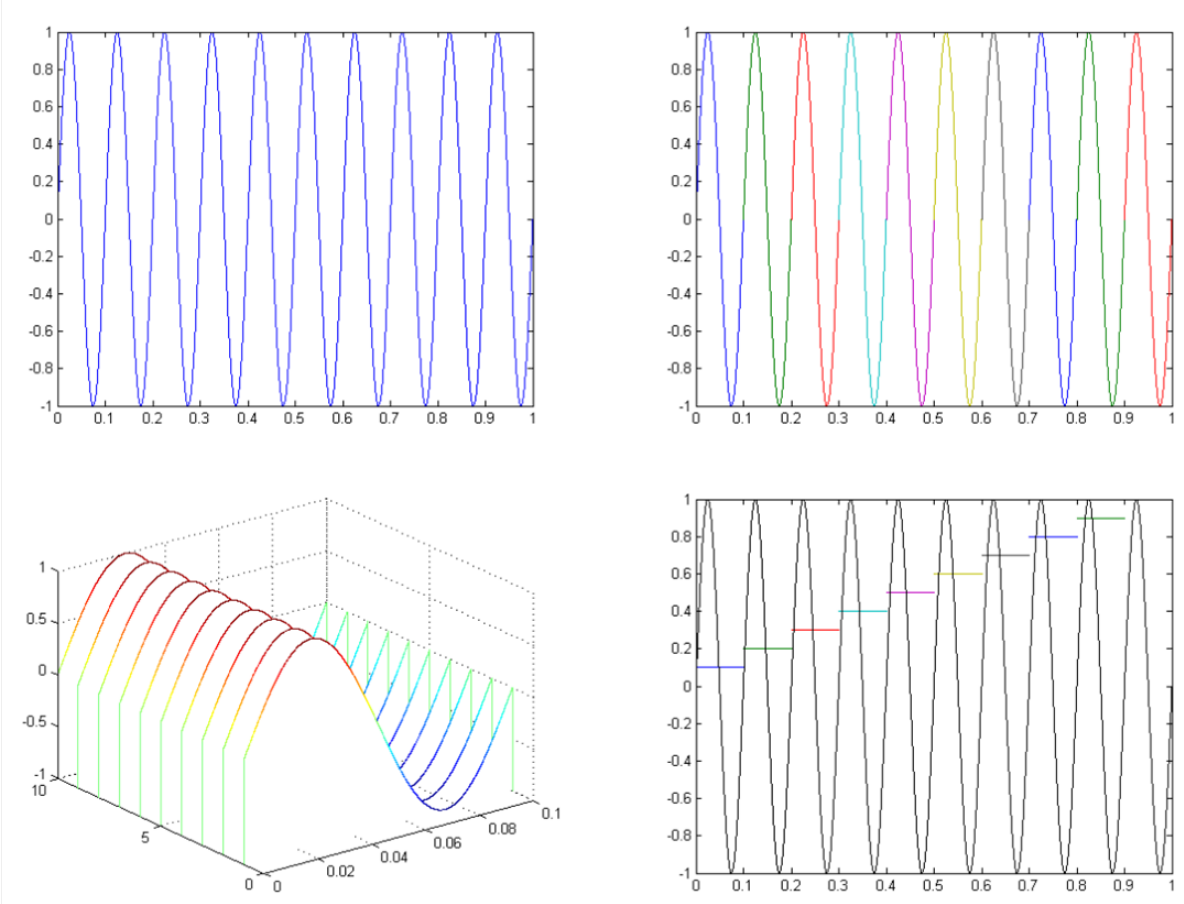

Figure 1.4: Découpage d'une sinusoïde en trames qui ne se chevauchent pas.

- T_K est la durée d'une trame.
- N_K est le nombre d'échantillons dans une trame, il est identique pour toutes les trames d'un même son.
- $k \in \{0 \dots K-1\}$ est un indice qui identifie la trame.
- $t_k^{(\text{de})}$ instant associé au premier échantillon de la trame k.
- $t_k^{(\text{fin})}$ instant associé au dernier échantillon de la trame k.
- $t_k^{(\text{ce})}$ instant associé au centre de la trame k.
- $t_{k,n}$ échelle de temps associée à la trame k donnant les instants associées aux échantillons $n = 0$ jusqu'à $n = N_K - 1$.

Du fait qu'il n'y a pas de chevauchement, on a

$$\boxed{T_K = \frac{T}{K} \text{ et } N_K = \frac{N}{K}} \tag{1.8}$$

D'une façon générale, le début, le centre et la fin d'une trame k sont définies par

$$\boxed{\text{Pour } k \in \{0 \dots K-1\}, \quad \begin{cases} t_k^{(\text{de})} = k\frac{T}{K} \\ t_k^{(\text{fin})} = t_k^{(\text{de})} + T_K - T_e \\ t_k^{(\text{ce})} = \frac{1}{2}\left(t_k^{(\text{de})} + t_k^{(\text{fin})}\right) \end{cases}} \tag{1.9}$$

Avec ces définitions et pour une trame k, on peut définir une échelle de temps et déterminer les valeurs du signal

$$\boxed{\begin{array}{c} \text{Pour } k \in \{0 \dots\} \text{ et } n \in \{0 \dots N_K - 1\} \\ \\ t_{k,n} = t_k^{(\text{de})} + nT_e \text{ et } x_{k,n} = x(t_{k,n}) = x\left[\frac{t_{k,n}}{T_e}\right] \end{array}} \tag{1.10}$$

Dans la dernière expression en bas à droite de l'équation (1.10), l'expression entre crochet est un indice

$$n' = \frac{t_{k,n}}{T_e} = k\frac{N}{K} + n$$

et cet indice est utilisé pour retrouver dans le signal x_n la valeur recherchée $x_{k,n} = x_{n'}$.

1.3.3 Découpage en trames avec chevauchement

Ètant donné un nombre total d'échantillons N, un nombre de trames K et du choix de chevauchement utilisé, on a les deux propriétés suivantes.

- Les instants de débuts des trames ne sont pas modifiées.

$$\boxed{t_k^{(\text{de})} = k\frac{T}{K}} \tag{1.11}$$

- La longueur des trames est augmenté d'un facteur $\frac{1}{1-\alpha}$.

$$\boxed{T_K = \frac{T}{K}\frac{1}{1-\alpha} \text{ et } N_K = \frac{N}{K}\frac{1}{1-\alpha}} \tag{1.12}$$

- Hormis la première trame, toutes les trames partagent avec la précédente αN_K échantillons. Et hormis la dernière trame, toutes les trames partagent avec la suivante αN_K échantillons.

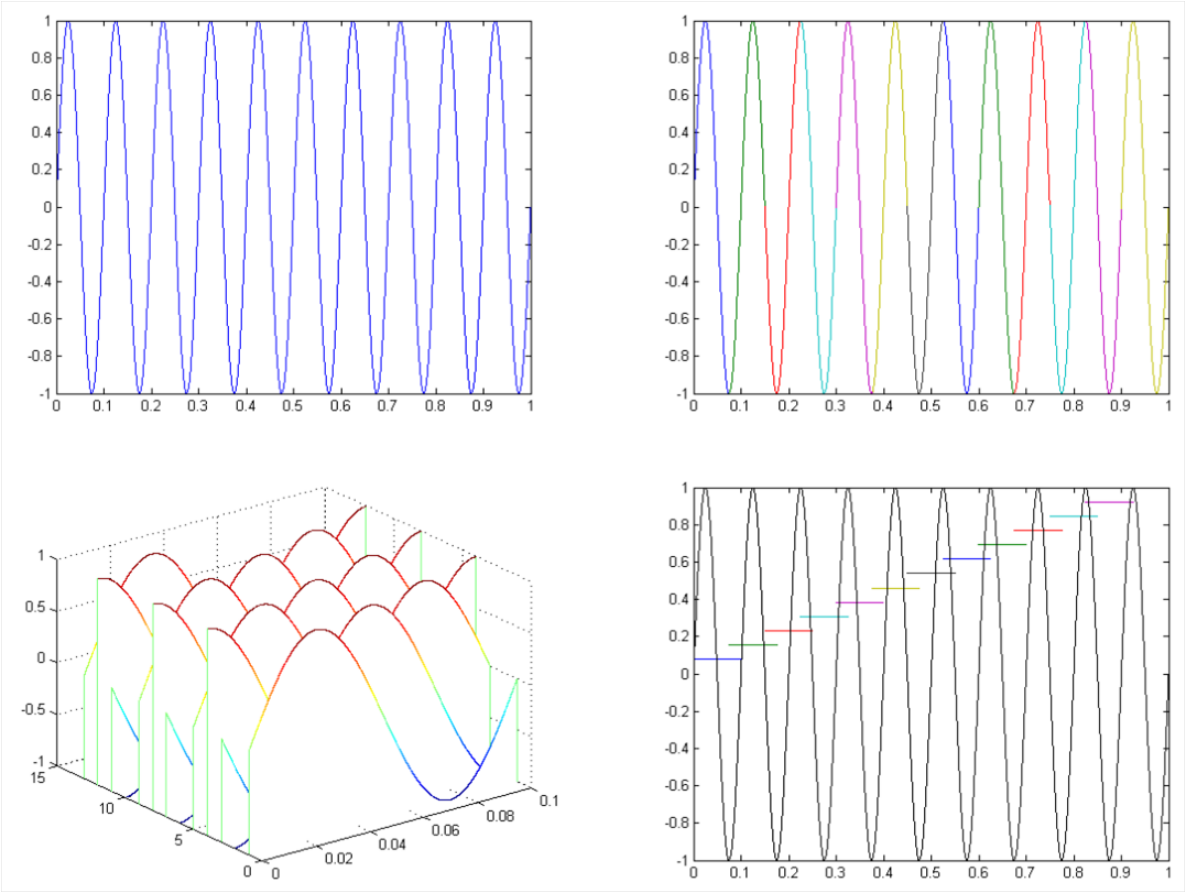

Figure 1.5: Découpage d'une sinusoïde en trames qui se chevauchent.

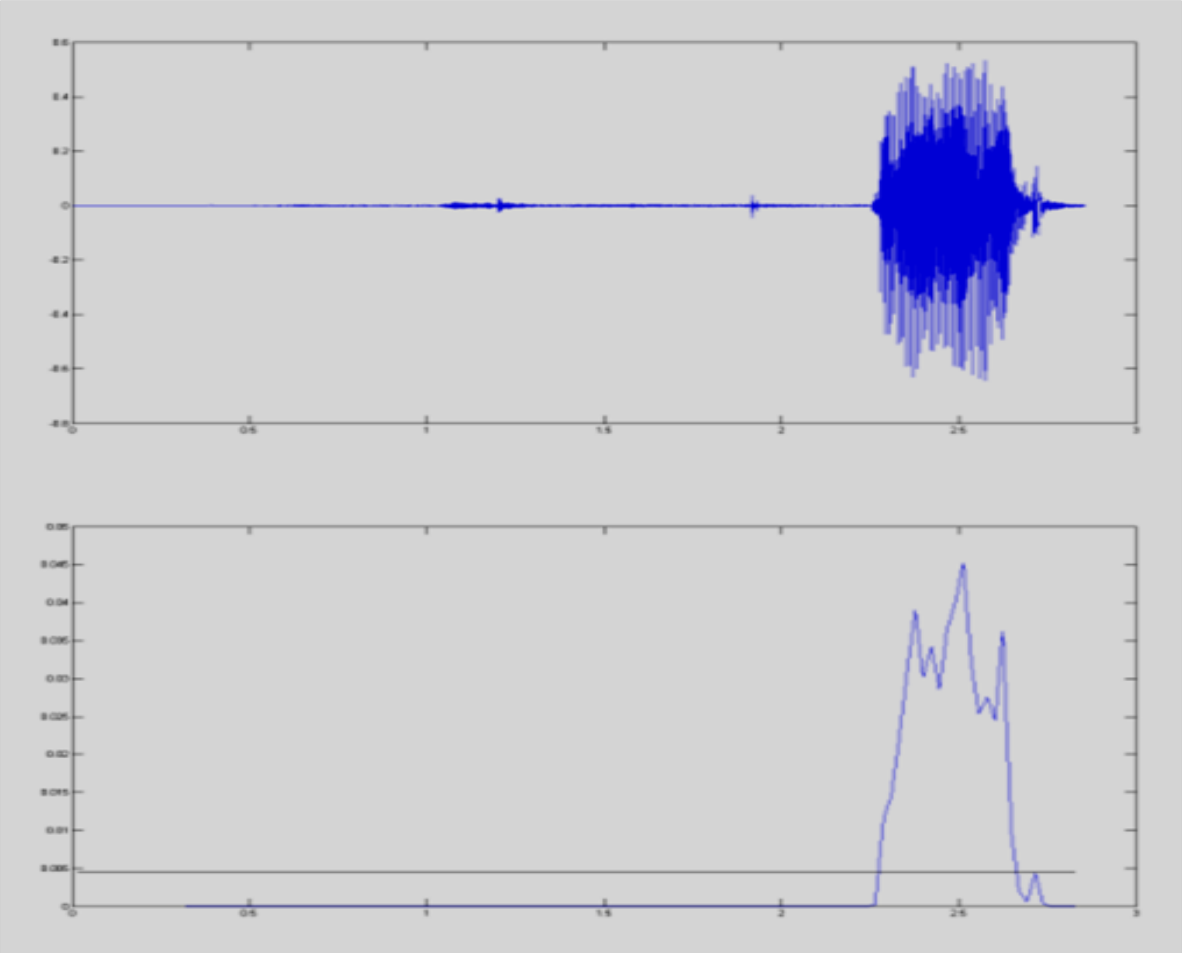

Figure 1.6: Signal temps discret représentant un signal sonore commençant et terminant par un silence.

La dernière affirmation se calcule ainsi pour $k \in \{0 \dots K-1\}$

$$t_k^{(\text{fin})} - t_{k+1}^{(\text{de})} + T_e = \left(k\frac{T}{K} + \frac{T}{K}\frac{1}{1-\alpha} - T_e\right) - (k+1)\frac{T}{K} + T_e = \frac{T}{K}\left(\frac{1}{1-\alpha} - 1\right) = \alpha T_K$$

Ainsi les équations (1.9) et (1.10) ne sont pas modifiées quand on met un chevauchement. En revanche l'équation (1.8) est remplacée par l'équation (1.12).

Lors de l'implémentation, ces formules ont été un peu modifiées pour tenir compte de ce que le découpage en trames ne tombe pas exactement sur un échantillon et qu'il faut donc commencer à partir de l'échantillon le plus proche.

1.4 Signaux sonores et silences

1.4.1 Puissance moyenne à court terme

En traitement du signal, on définit une puissance moyenne qui évolue dans le temps qui est appelée puissance moyenne à court terme. Cela se fait en introduisant une fenêtre notée w_T qui est un signal temps continu de durée T centré en zéro. La notion de fenêtre a déjà été présentée lors du cours sur la synthèse des filtres numérique à réponse impulsionnelle finie avec notamment la fenêtre rectangulaire et triangulaire. À temps continu, la fenêtre rectangulaire de largeur T est ainsi définie.

$$w_T(\tau) = \mathbf{1}_{\left[-\frac{T}{2},\frac{T}{2}\right]}(\tau)$$

La fenêtre triangulaire de largeur T est ainsi définie.

$$w'_T(\tau) = \left(1 - \frac{2t}{T}\right)\mathbf{1}_{\left[-\frac{T}{2},\frac{T}{2}\right]}(\tau)$$

Ici la fenêtre est utilisée, non pour modifier une réponse impulsionnelle mais un signal, ici $x^2(t)$. L'action de la fenêtre porte sur l'intervalle de temps centré en t. Aussi elle est d'abord centrée en t puis normalisée.

$$w_T(\tau) \mapsto w_T(t-\tau) \mapsto \frac{1}{\int_{\mathbb{R}} w_T(t-\tau)\,d\tau}\, w_T(t-\tau)$$

Ainsi modifiée la fenêtre est ensuite multipliée par $x^2(\tau)$

$$x^2(\tau) \mapsto x^2(\tau)\frac{1}{\int_{\mathbb{R}} w_T(t-\tau)\,d\tau}\, w_T(t-\tau)$$

Cette puissance moyenne à court terme, notée PMCT [4] [5] est

$$P(t) = \frac{\int_{\mathbb{R}} x^2(\tau) w_T(t-\tau)\,d\tau}{\int_{\mathbb{R}} w_T(\tau)\,d\tau} \tag{1.13}$$

Avec une fenêtre rectangulaire, $P(t)$ devient

$$\boxed{P(t) = \frac{1}{T}\int_{t-\frac{T}{2}}^{t-\frac{T}{2}} x^2(\tau)\,d\tau} \tag{1.14}$$

Ici le découpage en trame du signal se prête justement à un calcul très similaire.

- T est remplacée par la longueur de la trame T_K qui compte N_K échantillons.
- t est remplacée par $t_k^{(\text{ce})}$, le milieu de la trame k qui se trouve à peu près à l'échantillon $\frac{N_K}{2}$ de la trame k.

Cela permet ainsi d'approcher la discrétisation de l'équation (1.13).

$$\text{PMCT}_k = \frac{\sum_{n=0}^{N_K-1} x_k^2[n] w_{N_K}[n-\frac{N_K}{2}]}{\sum_{n=0}^{N_K-1} w_{N_K}[n-\frac{N_K}{2}]} \tag{1.15}$$

où $w_{N_K}[n]$ est une fenêtre de support $\{-\frac{N_K}{2}\ldots\frac{N_K}{2}\}$ **mais** avec des indices allant de 0 à N_K-1.

Avec une fenêtre rectangulaire,

$$\boxed{\text{PMCT}_k = \frac{1}{N_K}\sum_{n=0}^{N_K-1} x_k^2[n]} \tag{1.16}$$

1.4.2 Détection du début et de la fin d'un signal sonore.

La figure 1.6 montre un exemple de signal sonore commençant et terminant par un silence. On souhaite dans cet exemple retrouver l'instant à partir duquel le signal sonore commence et l'instant auquel ce signal sonore s'arrête.

On se donne un seuil, il est ici fixé à 10% de la valeur maximale.

$$\boxed{s_e = \frac{1}{10}\max_k(\text{PMCT}_k)} \tag{1.17}$$

Ce seuil est utilisé pour déterminer deux numéros de trames : $k^{(\text{de})}$ la première trame pour laquelle la puissance PMCT_k est supérieure s_e et $k^{(\text{fin})}$ la dernière pour laquelle ceci est vrai.

$$\boxed{k^{(\text{de})} = \min_k\{k|\,\text{PMCT}_k > s_e\} \text{ et } k^{(\text{fin})} = \max_k\{k|\,\text{PMCT}_k > s_e\}} \tag{1.18}$$

Les instants $t^{(\text{de})}$ et $t^{(\text{fin})}$ associés au début et à la fin du signal sonore sont les centres des trames $k^{(\text{de})}$ et $k^{(\text{fin})}$.

$$\boxed{\begin{cases} t^{(\text{de})} = t_k^{(\text{ce})} \text{ avec } k = k^{(\text{de})} \\ t^{(\text{fin})} = t_k^{(\text{ce})} \text{ avec } k = k^{(\text{fin})} \end{cases}} \tag{1.19}$$

où $t_k^{(\text{ce})}$ désigne l'instant associé au centre de la trame k.

[4] PMCT est défini dans [?] p. 41-42.

[5] Une interprétation en terme d'énergie existe dans page 3, section 2.1 dans [?].

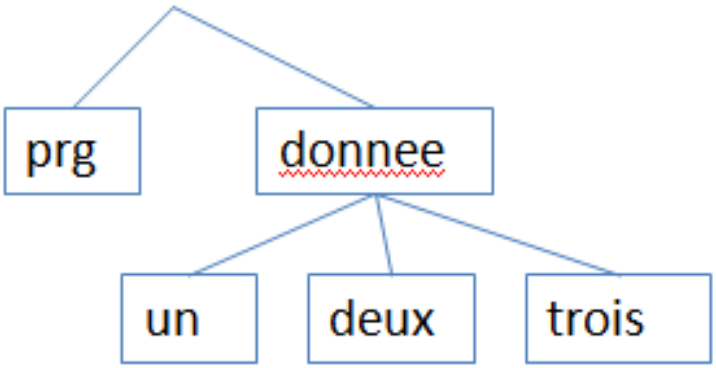

Figure 1.7: Organisation des données dans l'implémentation

1.4.3 Normalisation d'un signal sonore

La normalisation du signal a pour but de rendre comparable un son qui serait prononcé de façon forte et un son prononcé à voix plus basse. Cette normalisation consiste à imposer que la puissance moyenne soit fixe et en l'occurrence égale à 1, mais cette étape se fait après avoir retiré les silences du signal. [6]

Pour comparer les signaux, il est souhaitable que les signaux soient comparable en terme de puissance moyenne. Cette moyenne étant calculée sans considérer les silences. Pour obtenir que les signaux soient comparés, on effectue une *normalisation*, c'est-à-dire qu'on chacun des signaux par une valeur appropriée. Cette valeur est déterminée avec les étapes suivantes.

- On utilise la technique de détection du début et de fin du signal sonore présentée en section 1.4.2.
- On calcule la puissance moyenne en ne considérant que les trames qui se situent entre la première et la dernière des trames associées au signal sonore. On note $\mathbb{K}$ l'ensemble des trames où le son est plus fort que le seuil fixé par l'équation (1.17) et $|\mathbb{K}|$ le nombre de ces trames.

$$\boxed{P_{(\text{moy})} = \frac{1}{|\mathbb{K}|} \sum_{k \in \mathbb{K}} \text{PMCT}_k}$$

- La normalisation consiste à multiplier le signal par $\frac{1}{\sqrt{P_{(\text{moy})}}}$.

$$\boxed{x_n \mapsto \frac{1}{\sqrt{P_{(\text{moy})}}} x_n} \tag{1.20}$$

1.5 Indications sur l'implémentation proposée

Dans la simulation proposée, le répertoire principal est subdivisé en `prg` contenant les programmes Matlab (fichiers `.m` ou `.p`) et `donnee` qui contient les fichiers audio répartis en trois répertoires `un`, `deux` et `trois`. Cette organisation est schématisée sur la figure 1.7.

[6] Si on effectuait cette normalisation du signal avant de retirer les silences, un son avec beaucoup de silences que l'on aurait normalisé avec ses silences, aurait une puissance moyenne plus forte, une fois retirés les silences.

Chapter 2

Obtenir une classification grâce à des descripteurs

L'objectif au cours de la séance de TP correspondante est de classer les sons à partir de descripteurs. Un descripteur peut soit être une valeur pour l'ensemble du signal, soit un ensemble de valeurs, une pour chaque trame. On suppose dans cette partie qu'on dispose d'un certain nombre de descripteurs qui nous donnent des valeurs, une valeur par trame. La classification des sons est alors exposée en trois parties. La section 2.1 permet à partir de descripteurs d'évaluer numériquement une similarité entre deux sons. En fait on parle plutôt de distance entre sons, la distance correspondant ici à un défaut de similarité. une distance entre deux sons. La section 2.2 permet à partir de distances entre sons de prédire l'identité d'un son quelconque, c'est-à-dire s'il ressemble plus à un son de type `un`, `deux` ou `trois`. Enfin la section 2.3 montre comment on évalue généralement un classifieur.

2.1 Calcul d'une distance entre deux sons

La figure 2.1 schématise la façon de calculer une distance (i.e. une non-similarité) entre deux sons. La partie encadrée en pointillé figure la préparation des signaux avec les tâches suivantes.

- Chaque signal sonore est décrit avec $x_a[n]$ et $x_b[n]$ par un signal temps discret échantillonné à la même fréquence d'échantillonnage, mais pas forcément de même durée.
- Chaque signal est découpé en trames, de longueur et de chevauchement identique, le nombre de trames est a priori différent d'un signal à l'autre.
- Les trames silencieuses en début et fin de son sont retirées.
- Les valeurs des signaux au sein de chaque trames sont normalisées globalement au niveau du son, c'est-à-dire que si une trame est en moyenne deux fois plus forte qu'une autre trame du même son, elle reste deux fois plus forte après la normalisation. Par contre quand on compare deux sons, le fait de multiplier par deux toutes les valeurs d'un des deux sons, ne modifie pas la distance entre les deux sons.

On rajoute les notations suivantes :

- x_a et x_b désignent les sons a et b.
- $x_a[n]$ et $x_b[n]$ désignent les N_a et N_b valeurs des signaux associés aux sons a et b.
- $x_a[kn]$ et $x_b[kn]$ désignent les N_k valeurs des sons a et b pour la trame k, c'est l'équivalent pour le son a de ce qui était noté précédemment $x_k[n]$.
- J désigne le nombre total de descripteur.
- $j \in \{1 \dots J\}$ désigne un descripteur spécifique. PMCT est un exemple de descripteur, les autres seront décrits dans le chapitre 3.
- $z_a^{(j)}[k]$ et $z_b^{(j)}[k]$ désigne la valeur du descripteur j pour la trame k et les sons a et b.

Les calculs figurés par les flèches vertes joignant $x_a[kn]$ et $x_b[kn]$ à $z_a^{(j)}[k]$ et $z_b^{(j)}[k]$ correspondent aux calculs des valeurs des descripteurs, Les flèches rouges joignant $z_a^{(j)}[k]$ et $z_b^{(j)}[k]$ à $d(x_a, x_b)$ figurent les calculs permettant

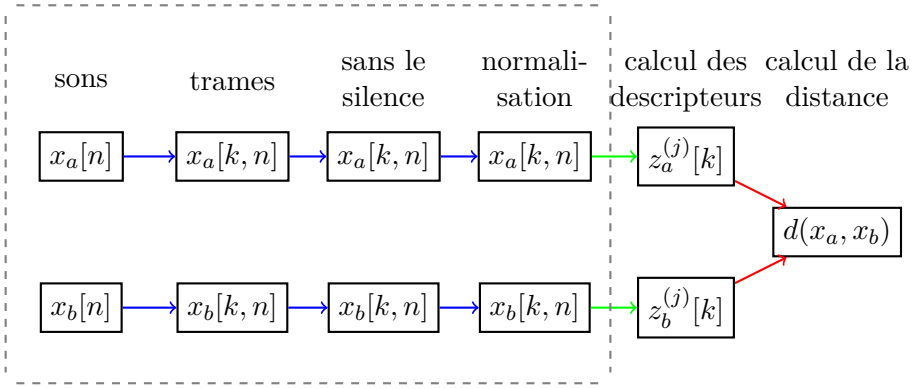

Figure 2.1: Schéma décrivant la mise en oeuvre du calcul de la distance entre deux sons noté x et x'. La partie encadrée en pointillé est la préparation des signaux transformant $x_a[n]$ et $x_b[n]$ en $x_a[k,n]$ et $x_b[k,n]$, n étant dans $x_a[n]$ et $x_b[n]$ l'indice de l'échantillon pour l'ensemble du son et pour $x_a[k,n]$ et $x_b[k,n]$, l'indice dans la trame k. La lettre j désigne un type de descripteur et celui-ci est calculé pour les deux sons et pour chaque $d(x_a,x_b)$ désigne la valeur de la distance entre les deux sons.

d'évaluer la distance à partir des valeurs des descripteurs, trois techniques sont proposées dans cette section.

2.1.1 Les descripteurs par trames sont transformés en descripteurs par son

On rajoute ici un autre sens au mot descripteur. Précédemment, il s'agissait d'un descripteur par trame, ici il s'agit d'un descripteur pour l'ensemble d'un son. Ce descripteur est identifié par j correspondant aux valeurs du descripteur par trame qui sont résumés dans une seule valeur. Ce descripteur est aussi identifié par la façon de faire ce résumé : μ pour la moyenne, σ pour l'écart-type, γ_1 et γ_2 pour les coefficients d'asymétrie et d'aplatissement.

Ces quatre descripteurs de sons sont calculés à partir des valeurs d'un descripteur obtenu pour chaque trame d'un son.

- La moyenne d'un descripteur j est la moyenne des valeurs de $z_k^{(j)}$ pour chaque trame k.

$$\boxed{\mu^{(j)} = \frac{1}{K}\sum_{k=0}^{K-1} z_k^{(j)}} \tag{2.1}$$

- L'écart-type d'un descripteur j est la racine carré de la moyenne des variances.

$$\boxed{\sigma^{(j)} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(z_k^{(j)} - \mu^{(j)}\right)^2}} \tag{2.2}$$

- Le coefficient d'asymétrie d'un descripteur j est

$$\gamma_1^{(j)} = \frac{1}{(\sigma^{(j)})^3}\frac{1}{K}\sum_{k=1}^{K}\left(z_k^{(j)} - \mu^{(j)}\right)^3 \tag{2.3}$$

- Le coefficient d'aplatissement d'un descripteur j est

$$\gamma_2^{(j)} = \frac{1}{(\sigma^{(j)})^4}\frac{1}{K}\sum_{k=1}^{K}\left(z_k^{(j)} - \mu^{(j)}\right)^4 \tag{2.4}$$

Ces quatre descripteur de sons permettent de définir différentes distances entre sons suivant ce qui est pris en compte. Par exemple en prenant en compte la moyenne et l'écart-type des descripteur $j \in \{1 \dots J\}$, on a la distance

suivante :

$$\boxed{d(x_a, x_b) = \sqrt{\sum_{j=1}^{J} \left(\mu_a^{(j)} - \mu_b^{(j)}\right)^2 + \left(\sigma_a^{(j)} - \sigma_b^{(j)}\right)^2}} \tag{2.5}$$

2.1.2 La distance entre deux sons est obtenue en comparant trame par trame, les valeurs des descripteurs.

Les sons a et b n'ont pas a priori la même durée et donc pas le même nombre de trames. On note $\min(K_a, K_b)$ le nombre de trames du son le plus court. On définit la distance comme la moyenne des distances entre les descripteurs trame par trame. Cette distance est en fait la racine carré de la somme des carrés des différences. Chaque son est décrit par $J \times K_a$ et $J \times K_b$ valeurs, on arrive à la définition suivante.

$$\boxed{d(x_a, x_b) = \sqrt{\frac{1}{\min(K_a, K_b)} \sum_{k=0}^{\min(K_a,K_b)-1} \sum_{j=1}^{J} \left(z_a^{(j)}[k] - z_b^{(j)}[k]\right)^2}} \tag{2.6}$$

Un important défaut de cette technique est qu'elle amène à considérer un son comme différent, lorsqu'il est prononcé avec une vitesse différente ou simplement une vitesse qui varie différemment.

2.1.3 La distance entre deux sons est obtenue en cherchant la déformation temporelle minimisant la différence entre les deux sons.

Notion de déformation temporelle dynamique

La figure 2.2 représente un même signal sinusoïdal colorié en vert $x_v(t)$ et différentes déformations de ce signal sinusoïdal, $x_b(t)$, colorées en bleu.

Il est d'usage lorsqu'on évalue la distance entre deux courbes définies sur un même intervalle de temps $[0, T]$, de calculer la moyenne quadratique de la différence entre $x_v(t)$ et $x_b(t)$.

$$\boxed{d_1(x_v, x_b) = \sqrt{\frac{1}{T} \int_0^T \left(x_v(t) - x_b(t)\right)^2 \, dt}} \tag{2.7}$$

Cette évaluation est souvent choisie pour sa simplicité, elle a une interprétation pertinente lorsque le signal déformé $x_b(t)$ resulte d'un bruit blanc gaussien centré additif, car dans ce cas $d_1(x_v, x_b)$ est une estimation de l'écart-type de cette gaussienne. Le graphe en haut à gauche de la figure 2.2 montre en bleu un signal déformé par un bruit blanc additif, on y observe que parfois la courbe bleue est au dessus de la courbe verte et parfois elle est en dessous.

Il existe dans la littérature scientifique une technique permettant d'évaluer la distance avec les différences en hauteur entre les signaux mais avec les distances euclidienne entre les courbes, c'est-à-dire pour chaque point de la courbe $x_b(t)$, on utilise la distance entre ce point et le point le plus proche de $x_v(t)$. Cette évaluation est pertinente lorsqu'on considère que la déformation de $x_b(t)$ résulte d'une part d'un variation de la valeur du signal $x_b(t_n^b) - x_v(t_n^v)$ mais aussi une variation de la base de temps $t_n^b - t_n^v$. Le graphe en haut à droite est obtenu en ajoutant deux bruits blancs gaussiens centrés additifs indépendants sur les valeurs du signal et sur les instants associés à ces valeurs.

Le graphe en bas à gauche est obtenu en ne modifiant que les bases de temps t_n^b par rapport à t_n^v, il s'agit d'un bruit blanc gaussien centré additif. C'est-à-dire qu'on a l'égalité entre les valeurs des signaux x_b et x_v mais les valeurs ont lieu à des instants différents.

$$x_b(t_n^b) = x_v(t_n^v)$$

Sur ce graphe, la courbe bleue présente des boucles suggérant que le temps s'inverse parfois. Ceci n'est pas une modélisation réaliste.

Le graphe en bas à droite est obtenu aussi en ne modifiant que les bases de temps t_n^b par rapport à t_n^v, mais pour plus de réalisme, l'ordre des évènements est respecté, il n'y a donc plus de boucles.

$$t_n^v \leq t_{n'}^v \quad \Rightarrow \quad t_n^b \leq t_{n'}^b$$

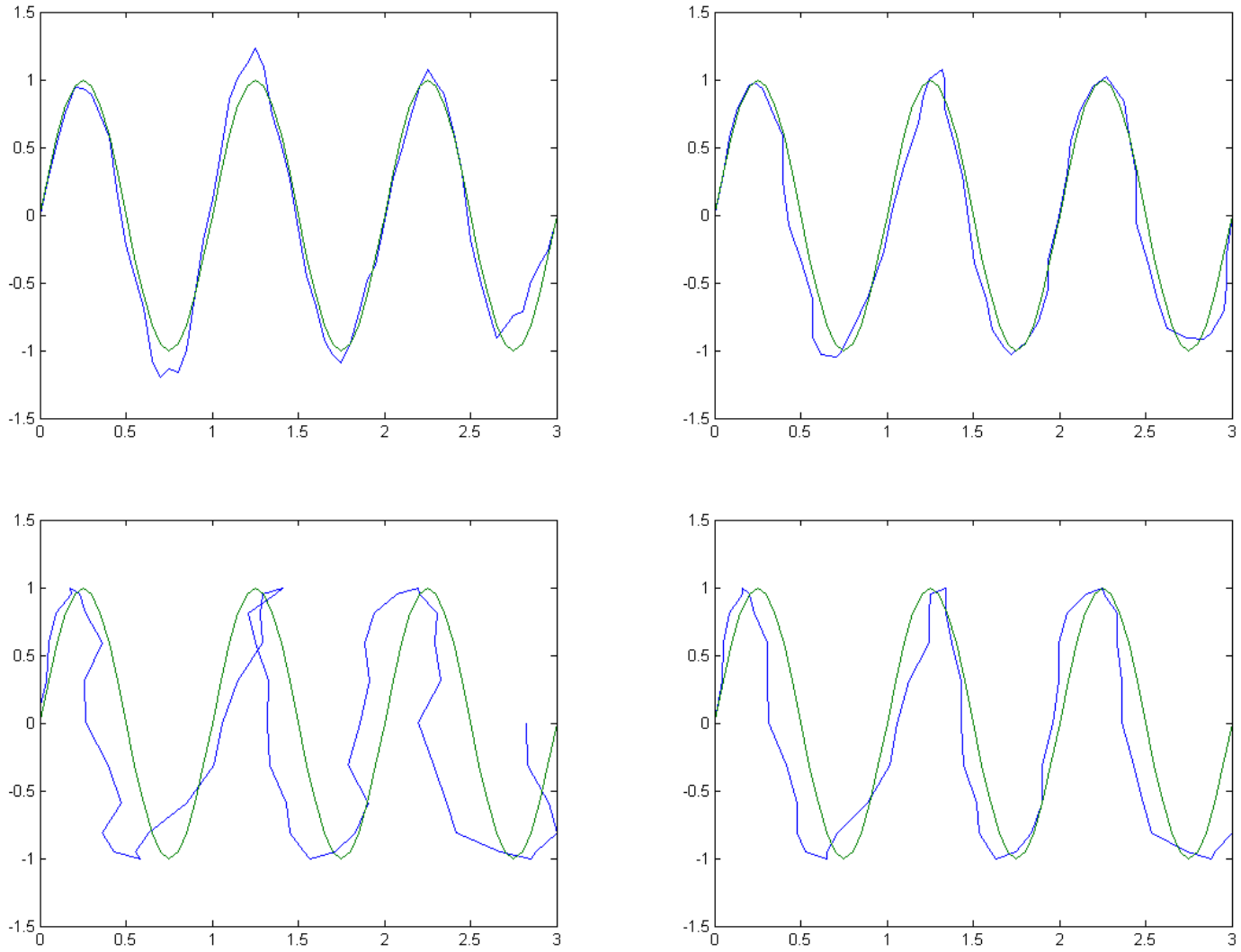

Figure 2.2: Quatre déformations possibles d'un même signal sinusoïdal représenté sur les quatre figures en vert. Le temps est représenté sur l'axe horizontal et les valeurs des signaux sont représentées sur l'axe vertical. En haut à gauche, le signal bleu est perturbé par un bruit additif, en haut à droite, il est perturbé par une bruit additif de moindre ampleur ajouté à la fois sur les valeurs du signal et celles du temps ; en bas à gauche, il est perturbé par un bruit additif ajouté seulement sur les valeurs de temps ; en bas à droite, le bruit additif ajouté seulement sur les valeurs de temps est corrigé par le fait qu'un évènement postérieur à un autre évènement doit rester postérieur après la perturbation.

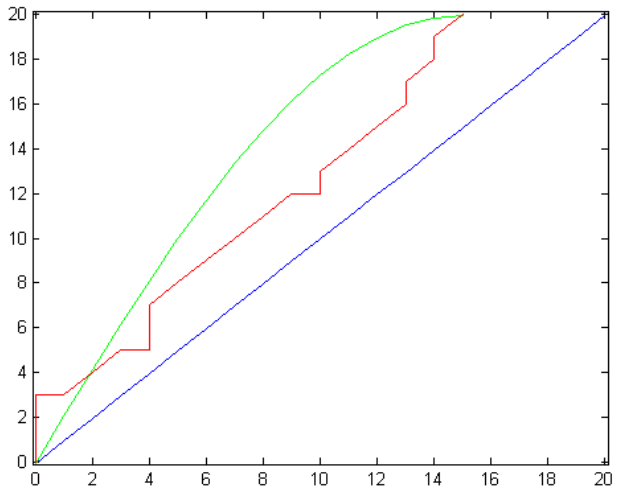

Figure 2.3: En vert et bleu, deux signaux sont définis sur des intervalles de temps différents (t_n, x_n), (t'_n, y_n). En rouge, le signal (t_n, x_n) est transformé par une table d'indexation en $(t'_{\Phi_y[n]}, x_{\Phi_x[n]})$ de façon à approcher le signal (t'_n, y_n).

Concrètement, les variations de t^b_n sont limitées par la contrainte que $t^b_n \leq t^b_{n+1}$ doit toujours être vrai. C'est ce type de déformation du signal que l'on cherche à prendre en compte. Quand on prononce un son plus lentement pour plus rapidement ou avec une vitesse qui varie au cours de la prononciation du son, il s'agit du même son. La distance devrait théoriquement ne pas être sensible à ce type de déformation. On appelle cette déformation, une déformation temporelle dynamique. [1]

Table d'indexation

La figure 2.3 illustre ce que l'on cherche à faire avec deux tables d'indexation. Le trait bleu associé à x_n est défini sur l'intervalle de temps $[0, 20]$, tandis que la courbe verte associée à y_n est définie sur l'intervalle de temps $[0, 15]$. Le signal $x_{\Phi_x[n]}$ est défini pour les entiers $\{0 \dots 21\}$ et est à valeurs parmi les valeurs de $\{x_0 \dots x_{20}\}$, ce sont des points qui sont sur la courbe rouge. La courbe rouge est en fait obtenue avec reliant les points $(t_{\Phi_y[n]}, x_{\Phi_x[n]})$.

$$
\begin{array}{llll}
\Phi_x[0] = 0 & \Phi_y[0] = 0 & (t_{\Phi_y[0]}, x_{\Phi_x[0]}) = (t_0, x_0) = (0, 0) & \text{position initiale en bas à gauche} \\
 & & \vdots & \\
\Phi_x[3] = 3 & \Phi_y[1] = 0 & (t_{\Phi_y[3]}, x_{\Phi_x[3]}) = (t_0, x_3) = (0, 3) & \text{déplacement vertical } \mathbf{V} \\
\Phi_x[4] = 3 & \Phi_y[4] = 1 & (t_{\Phi_y[4]}, x_{\Phi_x[4]}) = (t_1, x_3) = (1, 3) & \text{déplacement horizontal } \mathbf{H} \\
\Phi_x[5] = 4 & \Phi_y[5] = 2 & (t_{\Phi_y[5]}, x_{\Phi_x[5]}) = (t_2, x_4) = (2, 4) & \text{déplacement diagonal } \mathbf{D} \\
 & & \vdots &
\end{array}
$$

Cet exemple illustre le fait que deux tables d'indexation permettent de modéliser un déplacement horizontal et vertical.

En pratique, on n'utilise pas la base de temps, on modifie juste les indices avec les deux tables d'indexation de façon à minimiser le carré des différences entre $x_{\Phi_x[n]}$ et $y_{\Phi_y[n]}$.

$$y_{\Phi_y[n]} \approx x_{\Phi_x[n]} \tag{2.8}$$

On introduit de nouvelles notations.

- N_Φ: Les tables d'indexation sont définies sur une plage d'indices communs

$$n \in \{0 \dots N_\Phi - 1\}$$

[1] Cette notion est présentée en détail dans [**?**].

- N_x et N_y sont le nombre d'indices pour x_n et y_n. Chaque table d'indexation est à valeurs dans les indices sur lesquels x_n et y_n sont définis

$$\Phi_x[n] \in \{0 \dots N_x - 1\} \text{ et } \Phi_y[n] \in \{0 \dots N_y - 1\}$$

En pratique, on impose que le premier indice de Φ_x corresponde au point de départ de x_n, le dernier de Φ_x avec le point d'arrivée de x_n et de même pour Φ_y avec y_n.

$$\begin{cases} \Phi_x[0] = 0, & \Phi_x[N_\Phi - 1] = N_x - 1 \\ \Phi_y[0] = 0, & \Phi_y[N_\Phi - 1] = N_y - 1 \end{cases} \tag{2.9}$$

On introduit la contrainte modélisant le non-retournement du temps

$$\Phi_x[n] \leq \Phi_x[n+1] \text{ et } \Phi_y[n] \leq \Phi_y[n+1] \tag{2.10}$$

On rajoute une autre contrainte qui indique que la différence de temporalité entre les deux signaux ne doit pas être trop grande. Cette contrainte a aussi l'avantage de diminuer la complexité numérique de l'algorithme utilisé pour calculer cette nouvelle distance.

$$|\Phi_x[n] - \Phi_y[n]| \leq \delta \tag{2.11}$$

δ étant un paramètre à définir pour calculer cette distance.

Distance définie au moyen d'un problème d'optimisation

Pour chaque choix de Φ_x et Φ_y, on a une valeur de distance qui est obtenue classiquement avec la moyenne quadratique.

$$\boxed{d_{\Phi_x,\Phi_y}(x,y) = \sqrt{\frac{1}{N_x + N_y} \sum_{n=0}^{N_\Phi - 1} \left(x_{\Phi_x[n]} - y_{\Phi_y[n]}\right)^2}} \tag{2.12}$$

La distance prenant en compte de la déformation temporelle dynamique est alors défini en choisissant Φ_x et Φ_y de façon à respecter les contraintes définies par les équations (2.9), (2.10) et (2.11). et à minimiser $d_{\Phi_x,\Phi_y}(x,y)$.

$$\boxed{d_\delta(x,y) = \min_{\Phi_x,\Phi_y} d_{\Phi_x,\Phi_y}(x,y)} \tag{2.13}$$

Algorithme résolvant ce problème d'optimisation

La gauche de la figure 2.4 est similaire à la figure 2.3 avec en bleu et vert deux signaux (t_n, x_n) et (t_n, y_n) et en rouge, la déformation du premier signal pour se rapprocher du second : $(t_{\Phi_y[n]}, x_{\Phi_x[n]})$. La droite de cette figure montre les valeurs prises par $\Phi_x[n]$ et $\Phi_y[n]$ en représentant ces deux tables comme une succession de points formant un chemin reliant au départ $(0,0)$ au point d'arrivée $(N_x, N_y) = (5,3)$. Ces contraintes correspondent à l'équation (2.9). On satisfait la contrainte de l'équation (2.10) en imposant que le chemin soit composé seulement de déplacement verticaux, horizontaux ou diagonaux et en l'occurrence seulement d'un pas à la fois. La contrainte indiquée par l'équation (2.11) est indiquée par les traits noirs hachurés à l'extérieur de la zone autorisée sur la droite de la figure 2.4.

Les valeurs utilisées pour les signaux représentés sur la figure 2.4 sont

$$\begin{array}{lllllll} x_0 = 0 & x_1 = 1 & x_2 = 2 & x_3 = 3 & x_4 = 4 & x_5 = 5 \\ y_0 = 0 & y_1 = 2 & y_2 = 4 & y_3 = 5 \end{array}$$

Les tables d'indexations créées sont :

$$\begin{array}{llllll} \Phi_x[0] = 0 & \Phi_x[1] = 1 & \Phi_x[2] = 2 & \Phi_x[3] = 3 & \Phi_x[4] = 4 & \Phi_x[5] = 5 \\ \Phi_y[0] = 0 & \Phi_y[1] = 0 & \Phi_y[2] = 0 & \Phi_y[3] = 1 & \Phi_y[4] = 2 & \Phi_y[5] = 3 \end{array}$$

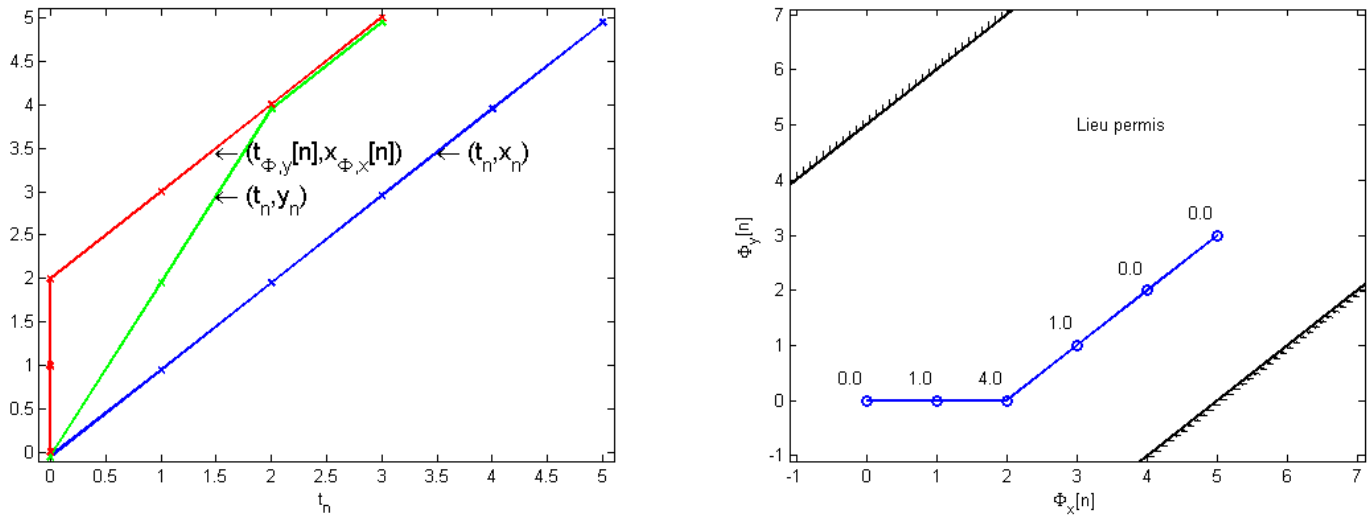

Figure 2.4: À gauche : représentation temporelle des signaux x_n, y_n et x_n transformés par déformation temporelle. À droite : représentation des tables d'indexation permettant de générer la déformation temporelle du signal x_n.

On appelle arête (*edge*) les traits associés au déplacement sur le graphe de droite. On appelle noeud (*node*, *vertice*) les points indiqués par des ronds bleus sur le graphe de droite. La localisation du point n est déterminé par $\Phi_x[n]$ et $\Phi_y[n]$, il se trouve que ces valeurs déterminent aussi

$$(x_{\Phi_x[n]} - y_{\Phi_y[n]})^2$$

Les nombres en bleus situés à côté de chaque noeud donnent la valeur de cette quantité. Ce sont les valeurs de x_n et y_n qui donnent à chaque point du graphe de droite une valeur pour $(x_{\Phi_x[n]} - y_{\Phi_y[n]})^2$ dépendant uniquement de la position du point. La distance entre les deux signaux telle que définie dans l'équation est à la somme des quantités inscrites à côté de chaque noeud. L'équation (2.13) indiquant que la distance est obtenue à partir du chemin minimisant $d_{\Phi_x,\Phi_y}(x,y)$ définie par l'équation (2.12). Il se trouve que le graphe à droite de la figure 2.4 donne une interprétation de cette minimisation. $d_{\Phi_x,\Phi_y}(x,y)$ est la somme des valeurs associées à chaque point et qui ne dépend que de la position des points. La minimisation consiste donc à trouver le chemin minisant la somme des valeurs de chaque noeud en utilisant que des déplacements verticaux, horizontaux ou diagonaux et menant d'un point de départ à un point d'arrivée tout en restant dans le domaine délimité par les zones hachurées.

L'algorithme permettant cette minimisation s'appelle l'algorithme de Dijkstra. Il consiste dans un premier temps à découvrir tous les noeuds possibles.

- Le premier noeud regardé est le point de départ et on lui affecte la valeur associée $(x_0 - y_0)^2$.
- On répète la tâche suivante

  1. Aller au voisin et lui donner la valeur de la distance entre le point de départ et ce voisin.
  2. On indique aussi à ce voisin le noeud précédent où l'on était.
  3. Si sur ce voisin il y a déjà une valeur, alors cette valeur doit être diminuée si la distance trouvée est moindre et on lui change l'origine que si l'on change sa valeur.
  4. Revenir au noeud précédent.
  5. Aller à l'étape 1 s'il y a un voisin disponible et non-exploré dans cette tâche.
  6. Revenir au premier voisin et s'autoriser à aller à une profondeur supplémentaire.

Lorsque ces tâches ne peuvent plus être répétés parce qu'on a été au maximum de la profondeur possible, alors on part du point d'arrivée et on remonte au point de départ en suivant les indications laissées à chaque noeud.

La figure 2.5 est assez similaire à la figure 2.4 mais avec d'autres valeurs pour les signaux x_n, y_n, $\Phi_x[n]$ et $\Phi_y[n]$.

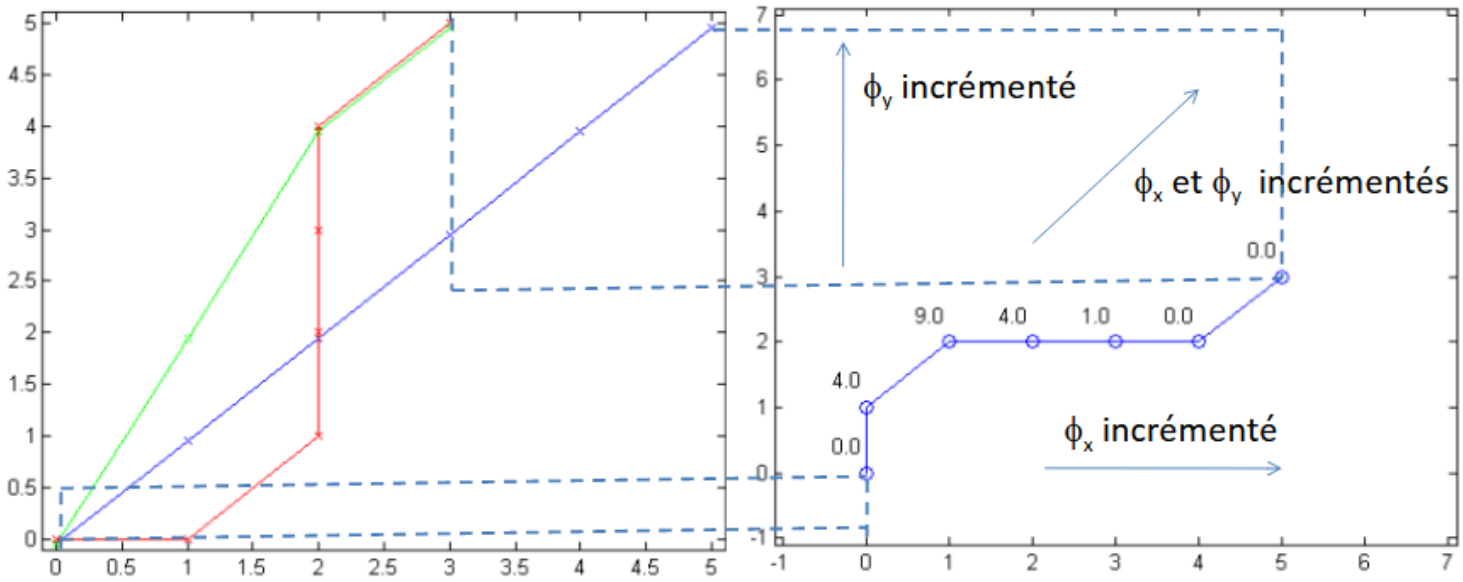

Figure 2.5: À gauche : représentation temporelle des signaux x_n, y_n et x_n transformés par déformation temporelle. À droite : représentation des tables d'indexation permettant de générer la déformation temporelle du signal x_n.

2.2 Prédiction par le plus proche voisin

2.2.1 Principe

On appelle **classifieur** le programme ou le procédé par lequel il est estimé que le son appartient à telle classe.

On appelle **prédiction** le fait pour un classifieur d'estimer qu'un son appartient à une classe en utilisant uniquement les données numériques disponible pour ce son.

La prédiction par le plus proche voisin est un cas particulier d'un classifieur appelé k-NN pour *Nearest Neighbor* ce qui signifie justement les k plus proche voisins. Ce classifieur est performant mais très coûteux en terme de complexité numérique dès lors que la base de donnée est assez importante. Il est aussi particulièrement simple à implémenter, ce qui est la raison principal de son choix ici. Il est utilisé ici dans le cas où $k = 1$.

Ce classifieur consiste pour chaque son inconnu appelé requête, de parcourir tous les sons de la base de données et pour chaque son de calculer sa distance avec la requête. Le classifieur détermine la classe comme étant celle du son ayant la plus petite distance avec la requête. Cette classe est appelée la prédiction faite par le classifieur pour cette requête. Il est intéressant de remarquer que ce choix de faire une classification en fonction d'un seul élément (ou de k éléments dans la version plus générale de cet algorithme) permet de s'adapter à une situation ou une classe regrouperait des éléments très hétérogènes.

La figure 2.6 montre comment une vingtaine de sons peuvent être visualisés sur un graphe. Chaque point correspond à un son, sa couleur (bleu pour `un`, vert pour `deux`, rouge pour `trois`) indique le mot que ce son vocalise. Cette information est ce que l'on cherche à prédire avec la prédiction par le plus proche voisin. La localisation de chaque point dépend des valeurs de deux indicateurs, μ_{PMCT} et σ_{PMCT}. Ces indicateurs ont été présentés dans les sections 1.4.1 et 2.1.1, c'est un programme informatique qui à partir des données enregistrées permet de calculer ces deux valeurs et ainsi de positionner ces points sur ce graphique. Pour illustrer le fonctionnement de cette technique de prédiction, voici son fonctionnement.

1. Sélectionnez dix points que vous numérotez de 1 à 10.
2. Considérer un autre point qui ne figure pas parmi ces points. Ce point est appelé requête.
3. Mesurez avec une règle la distance entre ce point requête et chacun de ces 10 points numérotés.
4. Regardez la couleur du point qui a la distance avec le point requête la plus faible (et qui est donc sur le graphique le plus proche du point requête parmi les 10 points numérotés).

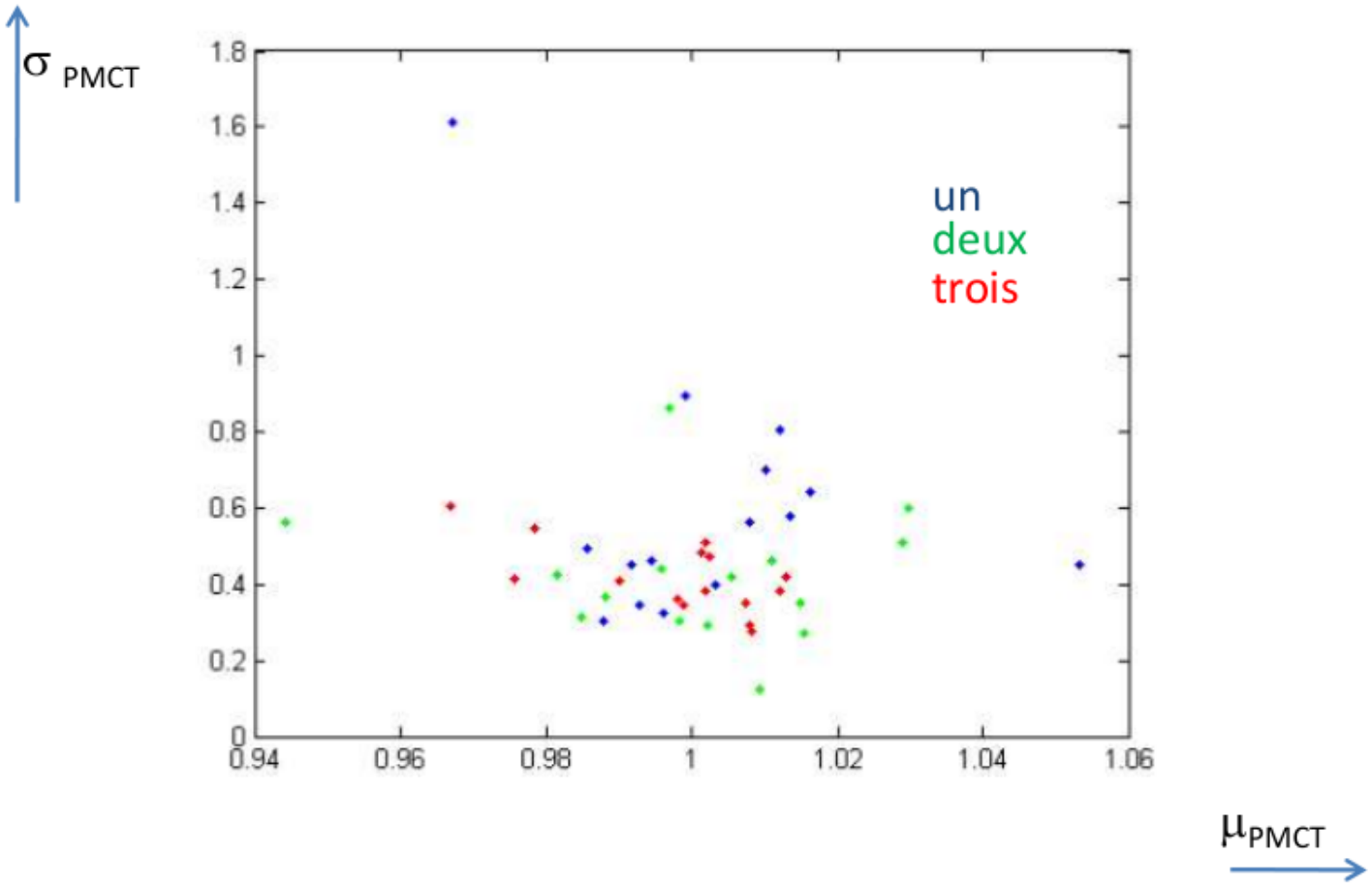

Figure 2.6: Visualisation de sons à travers les valeurs des indicateurs μ_{PMCT} et σ_{PMCT}

5. La prédiction consiste à attribuer au point requête la couleur du point le plus proche et par suite la classe, `un`, `deux` ou `trois` suivant la couleur observée.

2.2.2 Formalisation de la technique du plus proche voisin

Pour formaliser cette technique du plus proche voisin, on adopte les notations suivantes.

- x désigne en général un son défini pour plus de généralité par ses valeurs x_n.
- χ désigne un ensemble de sons.
- y désigne la classe à laquelle le son x appartient, (`un`, `deux` ou `trois`), est indiqué par son appartenance respective à χ_1, χ_2, χ_3.
- $\mathbf{d}(x_a, x_b)$ est une valeur positive correspondant à la distance entre les sons x_a et x_b. En fait pour plus de généralité, ce calcul repose sur la réalisation des étapes suivantes.
  - Découpage de x_a et x_b en trames décrit à la section 1.3.
  - Calcul d'un indicateur par trame pour x_a et x_b, illustré par PMCT décrite à la section 1.4.1.
  - Mise en oeuvre d'une des trois distance présentées dans les sections 2.1.1, 2.1.2, 2.1.3.
- $\mathbf{d}(x, \chi)$ désigne la distance entre le son x et un ensemble de sons regroupés dans χ. Du fait du choix de la technique du plus proche voisin, cette distance est ici définie par

$$\boxed{\mathbf{d}(x, \chi) = \min_{x' \in \chi} \mathbf{d}(x, x')} \tag{2.14}$$

  où $\min_{x' \in \chi} \mathbf{d}(x, x')$ désigne la valeur la plus faible parmi les distances mesurées entre un son donné x et tous les sons x' contenus dans χ.
- $\widehat{x}$ désigne la classe prédite. Cette classe est prédite de cette façon.

$$\boxed{\widehat{x} = \arg\min_{c \in \{1,2,3\}} \mathbf{d}(x, \chi_c)} \tag{2.15}$$

  où χ_1, χ_2, χ_3 regroupent des sons dont on sait qu'ils sont respectivement associés à `un`, `deux` et `trois`. et $\arg\min_{c \in \{1,2,3\}} \mathbf{d}(x, \chi_c)$ est ainsi défini.

$$\arg\min_{c \in \{1,2,3\}} \mathbf{d}(x, \chi_c) = \begin{cases} 1 & \text{si} \quad \mathbf{d}(x, \chi_1) \leq \mathbf{d}(x, \chi_2) \quad \text{et} \quad \mathbf{d}(x, \chi_1) \leq \mathbf{d}(x, \chi_3) \\ 2 & \text{si} \quad \mathbf{d}(x, \chi_2) \leq \mathbf{d}(x, \chi_1) \quad \text{et} \quad \mathbf{d}(x, \chi_2) \leq \mathbf{d}(x, \chi_3) \\ 3 & \text{si} \quad \mathbf{d}(x, \chi_3) \leq \mathbf{d}(x, \chi_1) \quad \text{et} \quad \mathbf{d}(x, \chi_3) \leq \mathbf{d}(x, \chi_2) \end{cases}$$

2.2.3 Illustration avec l'implémentation proposée en TP

La figure 2.7 illustre la façon dont l'algorithme est ici implémenté. Chaque bloc représente un répertoire contenant des fichiers. Les sons que l'on utilise pour tester et évaluer la performance d'un classifieur sont en réalité issus de la base de données. Il faudrait donc s'assurer qu'ils ne sont pas utilisés aussi pour l'apprentissage. La partie qui n'est pas encadré en pointillé n'est pas modifiée lors de la mise en oeuvre de la classification. Cette mise en oeuvre consiste à tirer aléatoirement une partie des sons des répertoires `un`, `deux` et `trois`, de les placer dans le répertoire `requête`, les autres sons sont placés dans trois nouveaux répertoires contenus dans le répertoire `simulation` et appelé `un`, `deux` et `trois`. L'information selon laquelle les sons placés dans `requête` n'est pas disponible sur les fichiers sons, mais elle est stockée pour être utilisée ultérieurement lors de l'évaluation de l'algorithme.

2.3 Évaluation de la performance d'un classifieur

En pratique quand on utilise un procédé de reconnaissance vocal, on a besoin de savoir dans quelle mesure on peut lui faire confiance. Cette mesure est ce qu'on appelle l'**évaluation de la performance**. Pour se faire, on utilise des sons dont on connaît la classe à laquelle ils appartiennent et on compare cette classe avec ce que le classifieur prédit. Il existe un très grand nombre de mesures de performance, on en présente trois types : la sensibilité globale, la précision et le rappel.

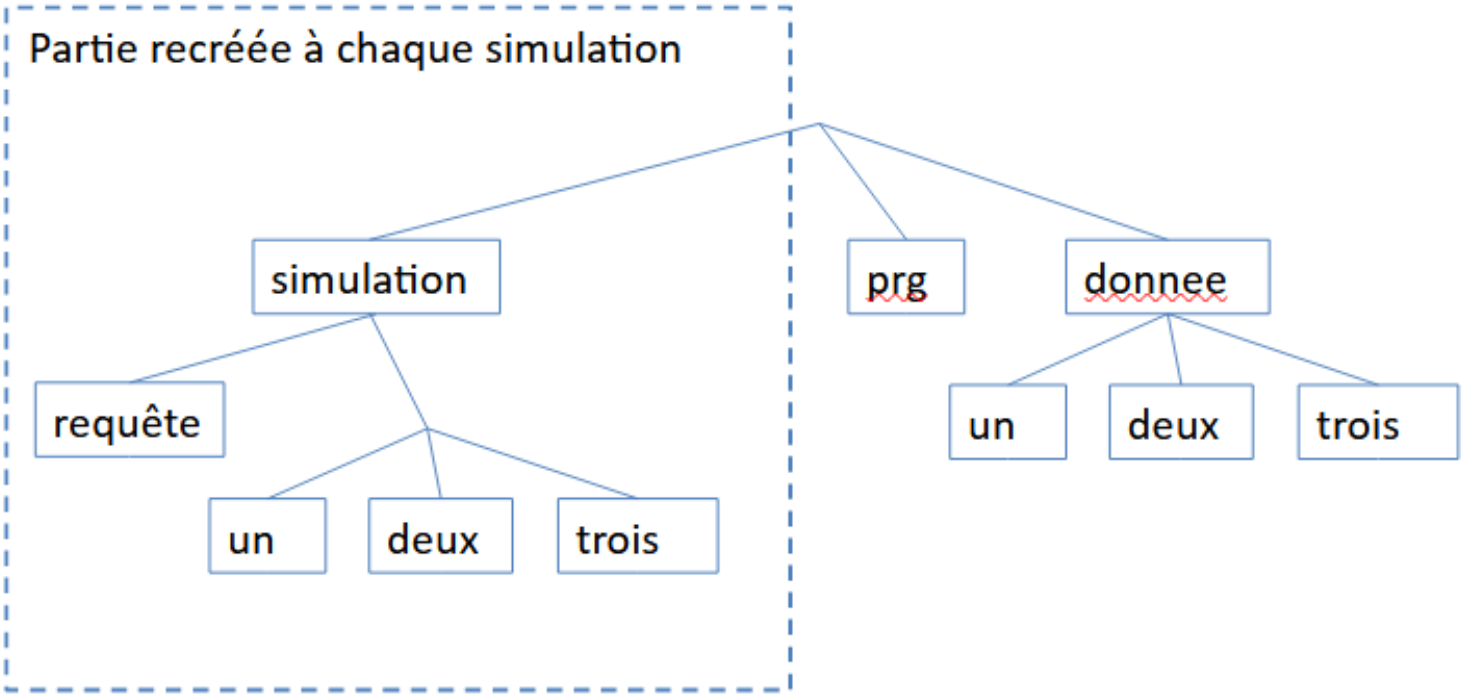

Figure 2.7: Schéma correspondant aux différents dossiers existant et créés pour réaliser la prédiction en utilisant la technique du plus proche voisin.

2.3.1 Matrice de confusion

Le rôle de la matrice de confusion est de résumer sous la forme d'une matrice les données utiles au calcul de l'évaluation de la performance. Ces données utiles consistent en le fait de connaître le nombre d'expérience où le son requête appartenait à telle classe et où la classe prédite est celle-là. Dans le cadre de cette étude, il s'agit d'une matrice 3×3 puisqu'il y a trois sons possibles.

- Chaque **ligne** est associée aux expérimentations sur les sons qui en réalité appartiennent à telle classe.
- Chaque **colonne** est associée aux expérimentations sur les sons qui ont été prédit comme étant de telle classe.
- La somme de toutes les cases est égale au nombre total d'expérimentations.
- À la case i, j, il y a le nombre d'expérimentations pour lesquelles le son appartient en réalité à la classe i et le classifieur l'a prédit comme étant de la classe j.
- m est un indice faisant référence à un son x.
- M est le nombre total de sons utilisés pour les différentes requêtes.
- $\mathbf{1}(\ldots)$ est la fonction caractéristique, elle prend la valeur 1 si les trois points de suspension sont remplacés par une affirmation exacte et par 0 si l'affirmation est fausse.
- $\mathbf{1}(y_m = c)$ vaut 1 si le son numéro m est de la classe c et $\mathbf{1}(y_m = c)$ vaut 0 si ce n'est pas le cas.
- $\mathbf{1}(\widehat{x}_m = c)$ vaut 1 si le son numéro m est prédit par le classifieur comme étant de classe c et $\mathbf{1}(\widehat{x}_m = c)$ vaut 0 si ce n'est pas le cas.
- $\mathbf{1}(y_m = i)\mathbf{1}(\widehat{x}_m = j)$ vaut 1 is le son numéro m est d'une part de la classe i et d'autre part a été prédit par le classifieur comme étant de la classe j.

Les composantes de la matrices de confusion sont :

$$\boxed{C_{ij} = \sum_{m=1}^{M} \mathbf{1}(y_m = i)\mathbf{1}(\widehat{x}_m = j)} \tag{2.16}$$

2.3.2 Sensibilité globale

La sensibilité globale, notée OA pour *Overall Assessment*, est le rapport entre le nombre de classifications correctes sur le nombre totales de requêtes, noté M. Les classifications correctes sont celles pour lesquelles $y_m = \widehat{x}_m$.

$$\boxed{\text{OA} = \frac{C_{11} + C_{22} + C_{33}}{3}} \tag{2.17}$$

2.3.3 Précision et rappel

La précision et le rappel sont des notions très utilisées pour les classifieurs binaires, c'est-à-dire pour lesquels, il n'y a que deux classes. Mais ces notions sont parfois étendues au cas plus général où on considère un nombre plus important de classes, c'est ce que nous utilisons ici.

Pour chaque classe c, on calcule deux indicateurs. La **précision** $\mathcal{P}_c$ est le rapport entre le nombre de requête correctement attribuées à la classe c et le nombre de requêtes que le classifieur évalué a prédit comme étant de la classe c. Le **rappel** $\mathcal{R}_c$ est le rapport entre le nombre de requête correctement attribué à la classe c et le nombre de requêtes qui dans la réalité correspondent à la classe c.

$$\boxed{\begin{cases} \mathcal{P}_1 = \frac{C_{11}}{C_{11}+C_{21}+C_{31}} \\ \mathcal{P}_2 = \frac{C_{22}}{C_{12}+C_{22}+C_{32}} \\ \mathcal{P}_3 = \frac{C_{33}}{C_{13}+C_{23}+C_{33}} \end{cases} \quad \text{et} \quad \begin{cases} \mathcal{R}_1 = \frac{C_{11}}{C_{11}+C_{12}+C_{13}} \\ \mathcal{R}_2 = \frac{C_{22}}{C_{21}+C_{22}+C_{23}} \\ \mathcal{R}_3 = \frac{C_{33}}{C_{31}+C_{32}+C_{33}} \end{cases}} \tag{2.18}$$

Ainsi la précision indique la probabilité qu'un échantillon, dont le classeur évalué indique qu'il s'agit de la classe i, soit effectivement de la classe i. Par contre, le rappel indique la probabilité pour qu'un échantillon dont on sait qu'il est de la classe i, soit détecté comme tel par le classifieur évalué.

2.3.4 Validation croisée

La **validation croisée** est une technique souvent utilisée pour la mise en oeuvre d'un algorithme de classification, elle n'est pas destinée à améliorer l'efficacité d'un classifieur, elle est destinée à évaluer plus précisément sa performance vis-à-vis d'une base de données en particulier lorsqu'il est compliqué d'augmenter le nombre de sons dont on connaît la classe à laquelle chacun appartient. La figure 2.8 illustre le fonctionnement de la validation croisée avec un schéma. En haut de cette figure se trouve 12 sons figurés par $x_1 \dots x_{12}$ et 12 valeurs notées $y_1 \dots y_{12}$ indiquant les véritables classes à laquelle chaque son appartient. Ici il s'agit d'une validation croisée en quatre blocs, ce qui amène à réaliser quatre apprentissages de classifieurs. `Exp1` désigne en haut la première expérience avec le premier apprentissage. Les trois premiers sons de la base d'apprentissage $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ servent de base pour tester le classifieur et sont affichés sur fond rouge. Le classifieur testé est entraîné en utilisant tous les autres sons $(x_4, y_4), \dots, (x_{12}, y_{12})$ qui sont eux affichés sur fond bleu. Le test du classifieur permet de calculer une matrice de confusion spécifique à `Exp1` et celle-ci permet de calculer la sensibilité globale OA_1. Les trois autres expériences sont fait de la même façon, mais en utilisant successivement comme sons tests les sons $(x_4, y_4), (x_5, y_5), (x_6, y_6)$, puis $(x_4, y_4), (x_5, y_5), (x_6, y_6)$ et enfin $(x_4, y_4), (x_5, y_5), (x_{12}, y_{12})$.

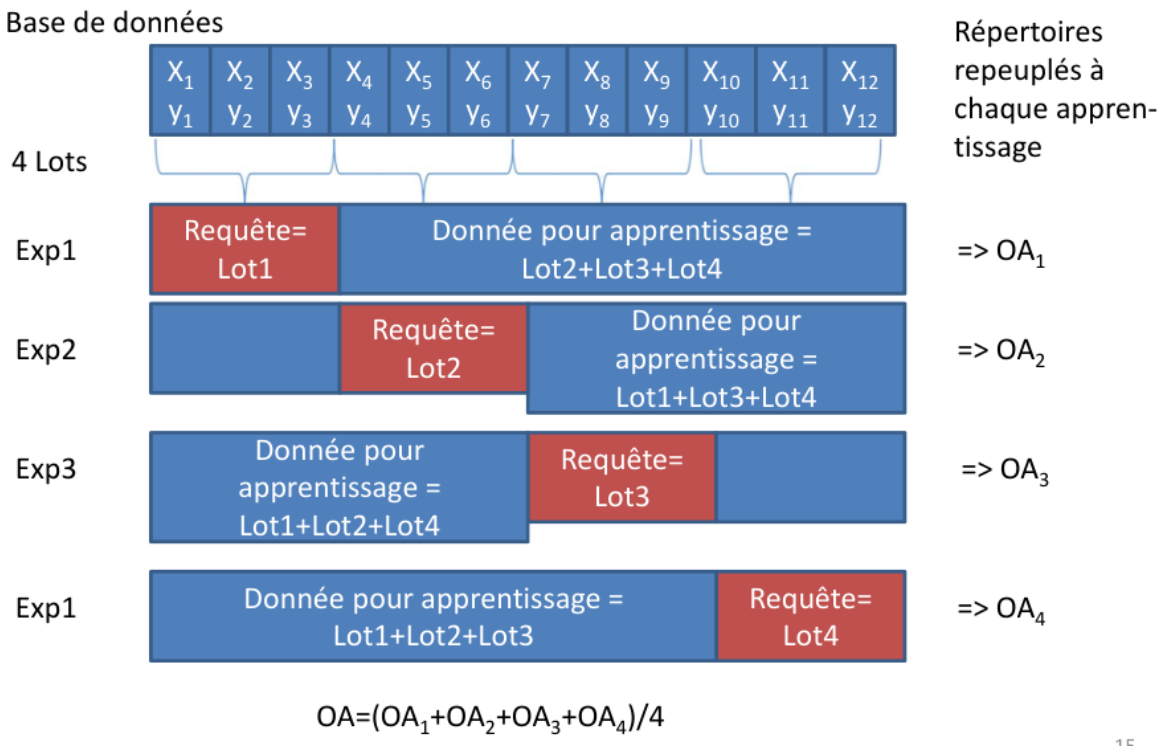

Figure 2.8: Schéma décrivant ce qu'est une validation croisée.

Chapter 3

Descripteurs de trames

La partie précédente ayant montrée comment un descripteur par trame permet de définir une technique de classification, cette partie traite de la détermination d'un descripteur par trame. L'inspiration qui a guidé l'élaboration des descripteurs provient de l'étude des caractéristiques des ondes sonores émises par un individu, c'est ce qui constitue la section 3.1 et qui montre l'intérêt de trouver la fréquence fondamentale et la plage de fréquences présentes dans le son émis. La section 3.2 présente différentes heuristiques permettant de décrire ce contenu spectral en indiquant la fréquence fondamentale, la fréquence moyenne et la largeur de spectre. La section 3.3 présente une technique plus systématique pour décrire le contenu spectral, c'est en utilisant un banc de filtres, chaque filtre constitue une heuristique donnant une information partielle sur le contenu spectral du son.

3.1 Modélisation de la production de la parole

La présentation de la modélisation de la façon dont un humain produit un son est faite en trois parties. La section 3.1.1 décrit le phénomène physique permettant de transformer une source sonore en un son dont le caractère plus ou moins aigü du son peut être déterminé assez précisément au moyen d'une note. La section 3.1.2 décrit brièvement les organes humains intervenant pour la production d'un son. Enfin la section 3.1.3 présente la notion d'analyse temps-fréquence illustré avec la présentation d'un spectrogramme.

3.1.1 Onde stationnaire

On dit qu'une onde est stationnaire si en chaque point, on observe que la surpression évolue dans le temps comme une sinusoïde. Par exemple l'équation suivante décrit une évolution sinusoïdale dont l'amplitude crête à crête dépend de l'indice x et vaut $\cos(2\pi\frac{x}{\lambda_0})$.

$$P(x,t) = \Delta P_0 \cos(2\pi f_0 t)\cos(2\pi\frac{x}{\lambda_0})$$

où λ et f_0 sont Cette onde peut approximativement correspondre à un cas particulier de l'équation (1.1) lorsque le coefficient $\frac{1}{r}$ est supposé approximativement constant et r est remplacé par x. La trigonométrie nous indique que le produit de deux cosinus est une différence entre deux sinus, $\cos(a)\cos(b) = \frac{1}{2}\cos(a-b) + \frac{1}{2}\cos(a+b)$. Cela permet de voir qu'une onde stationnaire est la somme d'une onde se propageant dans le sens des x croissants et une autre dans le sens des x décroissants.

$$P(x,t) = \frac{\Delta P_0}{2}\cos\left(2\pi f_0(t-\frac{x}{c})\right) + \frac{\Delta P_0}{2}\cos\left(2\pi f_0(t+\frac{x}{c})\right) \tag{3.1}$$

où $c = \lambda_0 f_0$ est la vitesse de propagation du son dans l'air. Cette modélisation correspond à $g_1 = g_2$ dans l'équation (1.1).

La figure 3.1 illustre de manière simplifiée la façon dont la forme géométrique d'une caisse de résonance détermine une fréquence de résonance. Le haut de la figure schématise al caisse de résonance qui est un pavé. La coupe est un rectangle. La zone colorée en bleu schématise une onde stationnaire qui consiste en une variation spatiale et temporelle avec une relation entre la longueur d'onde et la fréquence de la vibration indiquée dans l'équation 1.3 ($\lambda = c/f$ où c est ici la vitesse de propagation du son dans l'air, approximativement 1km toutes les 3 secondes). Les trois dessins illustrent qu'avec une même forme géométrique (un pavé de longueur d), 3 longueurs d'onde sont montrés.

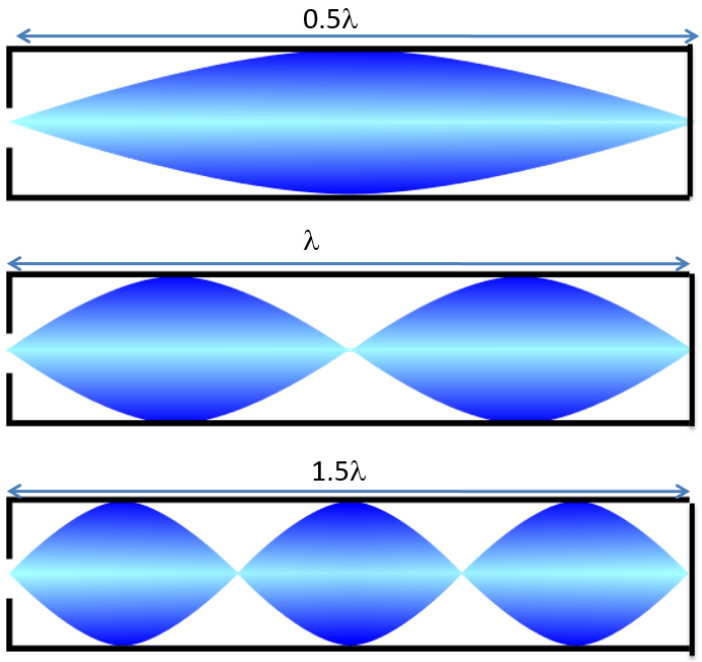

Figure 3.1: Pour une même taille donnée, un même obstacle peut être le siège de trois ondes stationnaires. La vitesse moyenne des particules d'air doit être nulle aux deux extrémités de la caisse de résonnance.

$\lambda = 2d$	$f = \frac{c}{2d}$	$d = \frac{\lambda}{2}$
$\lambda = d$	$f = 2\frac{c}{2d}$	$d = \lambda$
$\lambda = \frac{2}{3}d$	$f = 3\frac{c}{2d}$	$d = \frac{3\lambda}{2}$

Cette schématisation illustre un phénomène plus général, à savoir que l'émission d'un son à une certaine fréquence de base produit généralement des harmoniques qui sont des vibrations à des fréquences qui sont multiples de cette fréquence de base. La relation entre longueur du rectangle et longueur d'onde est ici déterminé par le fait que la vibration du son est nul sur la paroi (i.e. la paroi est fixe). Dans une onde stationnaire modélisée par l'équation (3.1), c'est à cet endroit que la pression est soit maximale soit minimale. L'annexe A formalise cette description de manière plus précise.

Une onde stationnaire peut aussi exister avec un rectangle ouvert sur un côté. Dans ce cas c'est la pression qui est fixée et la vibration qui est maximale à l'endroit où le rectangle est ouvert. Les relations entre longueurs d'onde, fréquences et longueur du rectangle sont alors modifiées.

$\lambda = 4d$	$f = \frac{c}{4d}$	$d = \frac{\lambda}{4}$
$\lambda = \frac{4}{3}d$	$f = 3\frac{c}{4d}$	$d = \frac{3\lambda}{4}$
$\lambda = \frac{4}{5}d$	$f = 5\frac{c}{4d}$	$d = \frac{5\lambda}{4}$

3.1.2 Modélisation de l'appareil phonatoire

On appelle appareil phonatoire l'ensemble des organes humains permettant de produire un son. L'objet de cette partie est une brève description de son fonctionnement.

La figure 3.2 représente une coupe de la tête d'une personne, (coupe dite sagittale), avec la gorge en bas à droite, la bouche au milieu à gauche, les dents le long d'une ligne horizontale placée légèrement au dessus de la bouche. Cette figure permet de modéliser la façon dont un son est formé.

- Le son est produit par la vibration d'un organe musculaire appelé corde vocale ou plis vocaux. Cet organe musculaire permet un contrôle de la fréquence du son émis à ce niveau.
- L'intensité du son est modulé par l'intensité du flux d'air sortant du poumon situé plus bas.

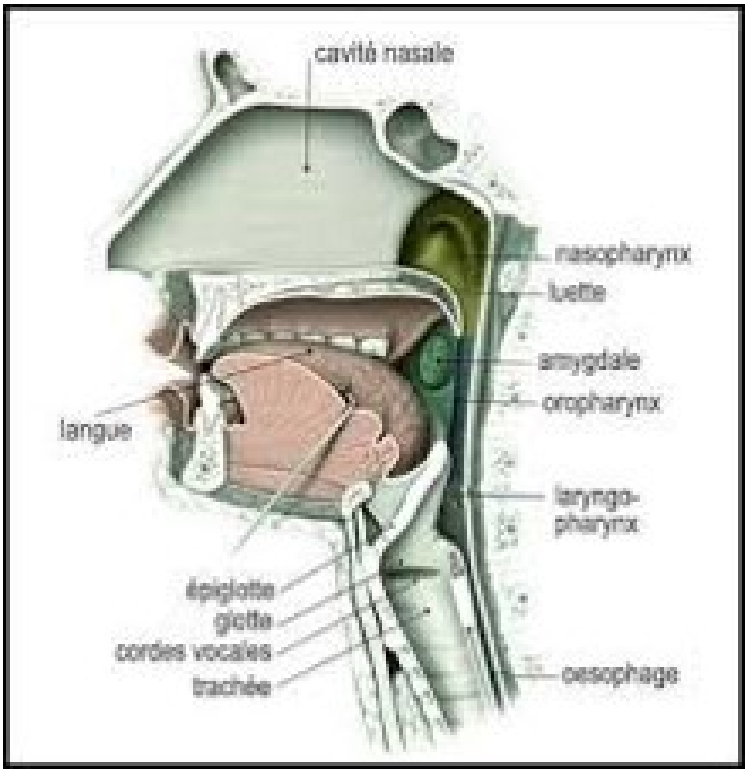

Figure 3.2: Diagramme représentant les organes de la parole

- Une première caisse de résonance est localisée dans la bouche et sa forme géométrique est modulée notamment par la position de la langue et l'ouverture de la bouche. Cette forme géométrique entraîne la formation de vibrations.
- Une deuxième caisse de résonance est localisée dans un espace libre au dessus de la bouche au niveau du nez, espace appelé cavité nasale. Cette deuxième caisse de résonance entraîne la formation d'un son avec une certaine fréquence.

Ainsi l'anatomie du corps humain permet de modifier l'intensité d'un son et la différence entre deux fréquences de résonance.

3.1.3 Spectrogramme et notion de description temps-fréquence

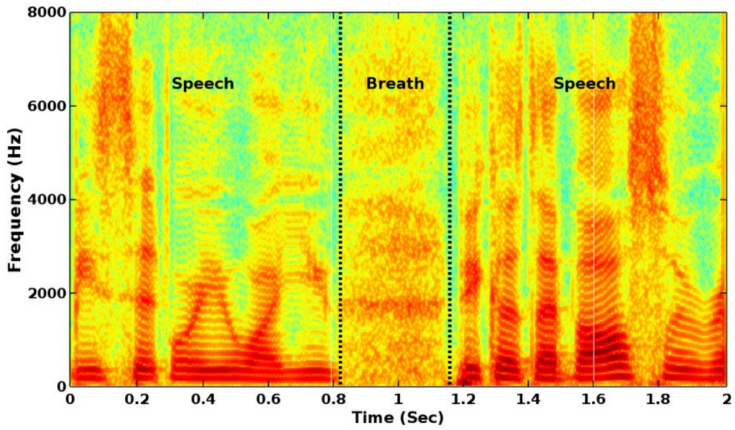

Figure 3.3: spectrogramme d'un signal sonore.

La figure 3.3 est un exemple de spectrogramme, c'est une visualisation d'un son, ici constitué d'une partie parlée à

gauche, un silence au milieu, et de nouveau une partie parlée à droite. L'axe horizontal est le temps en seconde. L'axe vertical est la fréquence en Hertz. Il s'agit d'une représentation temps-fréquence, ce qui peut paraître contradictoire dans le cadre de ce cours. En effet une description temporelle d'un signal ne dépend pas de la fréquence et une description fréquentielle d'un signal au moyen de son spectre ne dépend pas du temps. En fait la position latérale d'un point détermine une portion du signal sur un intervalle de temps. Tous les points d'une même verticale concernent la même portion du signal. Cette portion de signal est transformée en un spectre qui permet d'indiquer comment l'énergie de cette portion du signal est décomposée en une somme de signaux sinusoïdaux. Chaque point de cette verticale représente une de ces sinusoïdes, c'est la hauteur de ce point qui en déterminant une fréquence détermine aussi la sinusoïde. Et un point est d'autant plus sombre ou rouge que l'amplitude du signal sinusoïdal est plus importante. Sur ce spectrogramme, on observe une alternance de bandes sombres/rouges horizontales qui sont à la fois régulièrement espacées et avec des intensités qui diminues lorsque la fréquence augmente. Ce phénomène est lié au caisses de résonance qui produisent des sons avec une fréquence de base et des harmoniques qui sont des multiples de la fréquence de base.

Historiquement la recherche en reconnaissance de la parole s'est d'abord donnée comme guide le fait de modéliser les organes qui produisent ou reconnaissent la parole. En pratique la complexité numérique et la difficulté à modéliser limitaient la précision des modèles obtenues et par suite la performance. Il semble aujourd'hui admis que les techniques de machine learning permettent plus facilement d'obtenir des techniques efficaces. Il reste cependant de toutes ces études l'importance apportée à l'utilisation d'un certain type de descripteurs et par exemple l'importance d'estimer la fréquence fondamentale. L'utilité de s'intéresser à la fréquence fondamentale pour reconnaître une voyelle est expliquée en détail dans [**?**].

3.2 Estimation spectrale

L'objet de cette partie est de définir une application transformant les valeurs sur une trame k en une valeur donnée, appelée valeur du descripteur. La section 3.2.1 présente des techniques permettant d'estimer la fréquence fondamentale. La section 3.2.2 presente d'autres techniques permettant d'estimer d'autre caractéristiques du spectre, par exemple sa largeur.

3.2.1 Estimation de la fréquence fondamentale

Dans cette partie, les deux indicateurs présentés ont pour objectif d'estimer la fréquence fondamentale. [1]
Une autre technique existe pour mesurer la fréquence fondamentale en s'appuyant sur l'autocorrélation. [2]

Zero Crossing Rate

La figure 3.4 montre l'évolution de deux signaux en fonction du temps. Le premier en trait plus foncé/bleu est la superposition de 2 sinusoïdes. Le deuxième en trait plus clair/vert est une sinusoïde. Le descripteur [3] ZCR détermine le nombre de fois où la courbe traverse l'axe horizontal par une unité de temps. Ici il y a deux traversées toutes les secondes.

Le nombre de passage par zéro n'est en fait pas une notion qui se prête à une formalisation utilisant le temps continu. Le nombre moyen de passages par zéro pour la trame k est donnée par

$$\boxed{Z_k = \frac{1}{N_K} \sum_{n=0}^{N_K-1} \mathbf{1}\left(\text{signe}(x_k[n+1]) \neq \text{signe}(x_k[n])\right)} \tag{3.2}$$

Le fait de normaliser par la longueur de la trame (N_K) fait que Z_k n'est pas augmenté lorsque les trames sont plus longues.

Ce nombre moyen de passage par zéro est utilisé pour estimer la fréquence fondamentale, c'est la fréquence de la sinusoïde qui aurait ce nombre moyen de passages par zéros.

$$\boxed{f_k^{(\text{ZCR})} = \frac{Z_k f_e}{2}} \tag{3.3}$$

[1] Pour un son musical assez simple, cette fréquence correspond au *pitch* défini section C.3, p. 215-217 de [**?**].
[2] Cette technique est brièvement présentée p. 45-46 dans [**?**].
[3] Cette technique appelée *zero crossing rate* est évoquée dans la section 2.4 p. 12 de [**?**], et définie dans [**?**] p. 43-44.

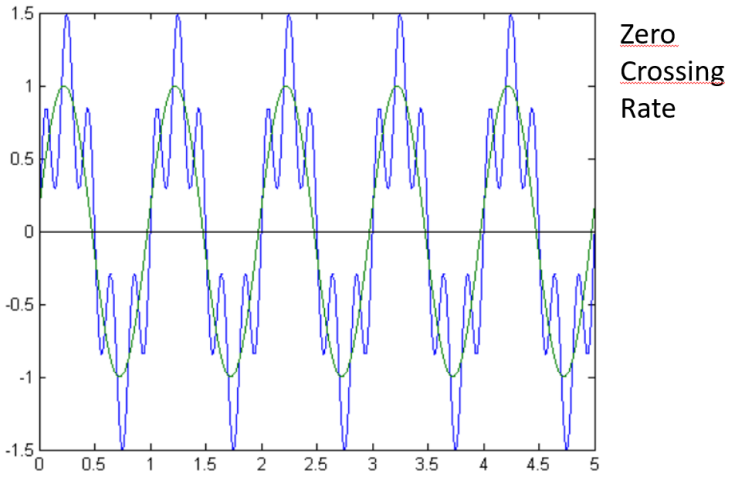

Figure 3.4: Graphe illustrant le fonctionnement de l'indicateur ZCR.

En effet, pour une sinusoïde de fréquence $f_k^{(\mathrm{ZCR})}$, sa période est $T = \frac{1}{f_k^{(\mathrm{ZCR})}}$ et le nombre moyen de passages par zéros est 2 divisé par le nombre de points sur une période Tf_e, aussi

$$f_k^{(\mathrm{ZCR})} = \frac{1}{T} = \left(\frac{2}{Tf_e}\right)\frac{f_e}{2} = \frac{Z_k f_e}{2}$$

Fréquence moyenne du spectre

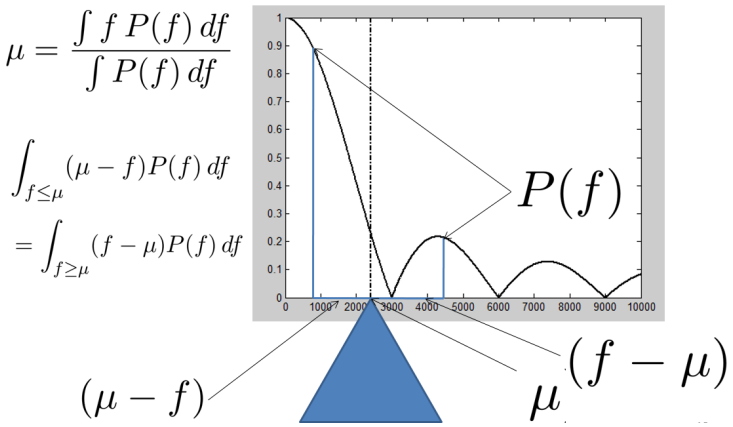

Figure 3.5: Illustration du calcul de la fréquence moyenne et de sa signification comme séparant en deux parties égales la surface sous la courbe.

La notion de temps-fréquence évoquée lors de la description du spectrogramme dans la section 3.1.3 permet de définir une densité spectrale d'énergie qui dépend du temps. Ici cela se fait en appliquant une fenêtre rectangulaire avec l'intervalle $\left[t - \frac{T}{2}, t - \frac{T}{2}\right]$ sur le signal $x(t)$.

$$\boxed{\left|\widehat{X}_t(f)\right|^2 = \left|\mathrm{TF}\left[x(t)\mathbf{1}_{\left[t-\frac{T}{2},t-\frac{T}{2}\right]}(t)\right](f)\right|^2} \tag{3.4}$$

Ensuite cette densité spectrale dépendant du temps et de la fréquence est une utilisée comme pondération dans $\mu(t)$

qui est appelée la fréquence moyenne (son unité est bien le Hertz).

$$\boxed{\mu(t) = \frac{\int_0^{+\infty} f \left|\widehat{X}(f)\right|^2 df}{\int_0^{+\infty} \left|\widehat{X}(f)\right|^2 df}} \tag{3.5}$$

On peut voir ce calcul de moyenne comme assez similaire au calcul de l'espérance d'une variable aléatoire à partir de la densité de probabilité d'une variable aléatoire : $E[T] = \int_{\mathbb{R}} t f_T(t)\, dt$ où ici f jouerait le rôle de t et $\frac{\left|\widehat{X}(f)\right|^2}{\int_0^{+\infty} \left|\widehat{X}(f)\right|^2 df}$ jouerait le rôle de $f_T(t)$.

La figure 3.5 donne une justification de la définition de $\mu(t)$ dans l'équation (3.5). La courbe en noir notée $P(f)$, correspondant en réalité à $|\widehat{X}_t(f)|^2$, est la densité spectrale d'énergie d'une portion du signal $x(t)\mathbf{1}_{[t-\frac{T}{2},t+\frac{T}{2}]}(t)$. Le trait semi-pointillé vertical est positionné

Cette notion est adaptée aux signaux à temps discret en considérant la portion du signal [4] sur une trame $x_k[n]$ et en lui calculant la transformée de Fourier discrète considérant que la longueur de la trame est suffisante pour que cette transformée de Fourier discrète soit une bonne approximation de la transformée de Fourier de cette portion de signal. On utilise ici un indice l pour les fréquences et on ne considère que la première moitié des N_K indices parce que la deuxième moitié des indices correspond aux fréquences négatives qui n'est pas prise en compte dans l'équation (3.5). Cette première moitié des indices est notée

$$\boxed{N_K^{|2} = \left\lfloor \frac{N_K}{2} \right\rfloor} \tag{3.6}$$

où $\lfloor \dots \rfloor$ désigne l'arrondi par en dessous. Les $N_K^{|2}$ premières fréquences sont $\left[0 \dots \frac{N_K^{|2} f_e}{N_K}\right]$. Ces différentes valeurs sont écrites au moyen d'un nouvel indice l dont dépend la fréquence.

$$\boxed{f_l = \frac{l f_e}{N_K} \text{ pour } l \in \{0 \dots N_K^{|2}} \tag{3.7}$$

La pondération utilisée est la densité spectrale de puissance de la trame k.

$$\boxed{\left|\widehat{X}_k[l]\right|^2 = \left|\text{TFD}\left[x_k[n]\right][l]\right|^2} \tag{3.8}$$

Un estimateur de la fréquence moyenne du spectre μ_k [5] est

$$\boxed{\mu_k = \frac{\sum_{l=0}^{N_K^{|2}} f_l \left|\widehat{X}_k[l]\right|^2}{\sum_{l'=0}^{N_K^{|2}} \left|\widehat{X}_k[l']\right|^2}} \tag{3.9}$$

Un indice l' est utilisé parce qu'il n'y a en fait aucun lien entre le compteur utilisé dans la sommation au niveau du numérateur et le compteur utilisé dans la sommation au niveau du dénominateur.

3.2.2 Autre descripteurs du spectre

Largeur du spectre

La largeur du spectre est estimée connaissant $\mu(t)$ avec

$$\boxed{\sigma(t) = \frac{\sqrt{\int_{\mathbb{R}_+} (f - \mu(t))^2 \left|\widehat{X}_t(f)\right|^2 df}}{\sqrt{\int_{\mathbb{R}_+} \left|\widehat{X}_t(f)\right|^2 df}}} \tag{3.10}$$

[4] Cette application de la transformée de Fourier discrète sur le signal découpé peut être vu comme l'utilisation d'une transformée de Fourier à court terme, p. 99 de [**?**].

[5] Cet estimateur est défini dans la section 6.1.1 p. 13 de [**?**].

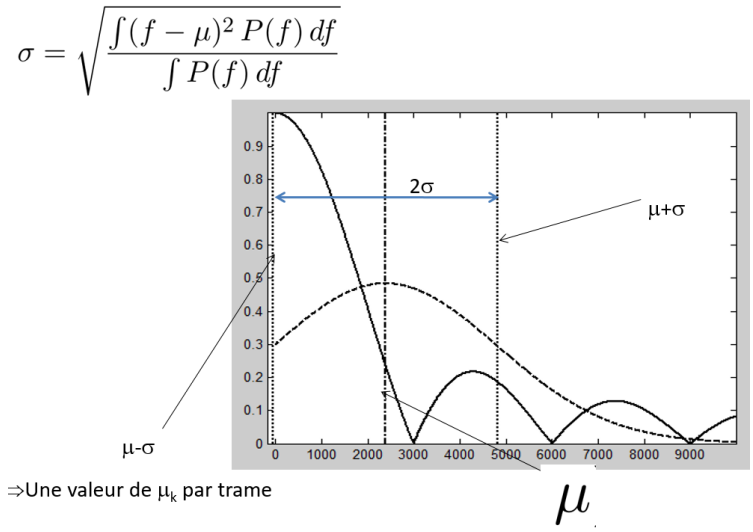

Figure 3.6: Illustration graphique sur un exemple du sens donné à la notion d'estimation de la largeur d'un spectre.

La figure 3.6 représente l'exemple d'un spectre d'une porte. La fréquence est en abscisse. Ce spectre a pour équation un sinus cardinal avec un demi-lobe principal de 3000Hz. Cette figure illustre deux descripteurs μ et σ. μ est la fréquence moyenne et σ est l'écart-type. Les deux sont calculés comme si les spectres étaient nuls pour les fréquences négatives (s'ils étaient calculés sur les vrais spectres, comme leur module est symétrique, la moyenne serait nulle. μ est indiqué par un trait vertical en pointillé. σ est représenté par deux traits verticaux autour de celui associé à μ. Ces deux traits verticaux indiquent que si le spectre était une gaussienne (courbe en traits pointillés), 66% de sa surface serait incluse dans ces deux traits verticaux en pointillés. Ce graphe n'est en fait pas tout à fait exact, d'une part la proportion de 66% et l'axe de symétrie au centre de la gaussienne n'est ici pas valable puisqu'on ne considère que les fréquences positives, d'autre part la moyenne et le calcul de la largeur se fait en considérant la densité spectrale et non le module de la transformée de Fourier et cette densité spectrale n'est donc pas exactement un sinus cardinal.

De même pour le calcul de la moyenne temporelle, l'équation (3.10) décrivant l'estimation de la largeur du spectre peut se voir comme l'estimation de l'écart-type d'une variable aléatoire positive de densité de probabilité $f_T(t)$

$$\sqrt{\mathrm{var}(T)} = \sqrt{E[(T - E[T])^2]} = \sqrt{\int_{\mathbb{R}_+} (t - E[T])^2 f_T(t)\, dt}$$

où ce qui remplace ici la variable t est f et la densité de probabilité $f_T(t)$ est

$$\frac{1}{\int_{\mathbb{R}_+} \left|\widehat{X}_t(f)\right|^2 df} \left|\widehat{X}_t(f)\right|^2$$

En utilisant les mêmes notations que pour (3.9), une estimation de la largeur du spectre [6] est

$$\sigma_k = \sqrt{\frac{\sum_{l=0}^{N_K^{\vert 2}} (f_l - \mu_k)^2 \left|\widehat{X}_k[l]\right|^2}{\sum_{l'=0}^{N_K^{\vert 2}} \left|\widehat{X}_k[l']\right|^2}} \tag{3.11}$$

[6] Cet estimateur est défini dans la section 6.1.2 p. 13 de [?].

Coefficient d'asymétrie du spectre

Un estimateur du coefficient d'asymétrie (*skewness*) [7] est donné par

$$\gamma_1(t) = \frac{1}{\sigma^3(t)} \frac{\int\limits_{\mathbb{R}_+} (f - \mu(t))^3 \left|\widehat{X}_t(f)\right|^2 df}{\int\limits_{\mathbb{R}_+} \left|\widehat{X}_t(f)\right|^2 df}$$

qui devient avec les notations utilisées dans (3.11).

$$\gamma_1[k] = \frac{1}{\sigma_k^3} \frac{\sum_{l=0}^{N_K^{\vert 2}} (f_l - \mu_k)^3 \left|\widehat{X}_k[l]\right|^2}{\sum_{l'=0}^{N_K^{\vert 2}} \left|\widehat{X}_k[l']\right|^2} \tag{3.12}$$

De même, un estimateur du coefficient d'aplatissement (*kurtosis*) [8] est donné par

$$\gamma_1(t) = \frac{1}{\sigma^4(t)} \frac{\int\limits_{\mathbb{R}_+} (f - \mu(t))^4 \left|\widehat{X}_t(f)\right|^2 df}{\int\limits_{\mathbb{R}_+} \left|\widehat{X}_t(f)\right|^2 df}$$

qui devient avec les notations utilisées dans (3.11).

$$\gamma_2[k] = \frac{1}{\sigma_k^4} \frac{\sum_{l=0}^{N_K^{\vert 2}} (f_l - \mu_k)^4 \left|\widehat{X}_k[l]\right|^2}{\sum_{l'=0}^{N_K^{\vert 2}} \left|\widehat{X}_k[l']\right|^2} \tag{3.13}$$

3.3 Utilisation de bancs de filtres

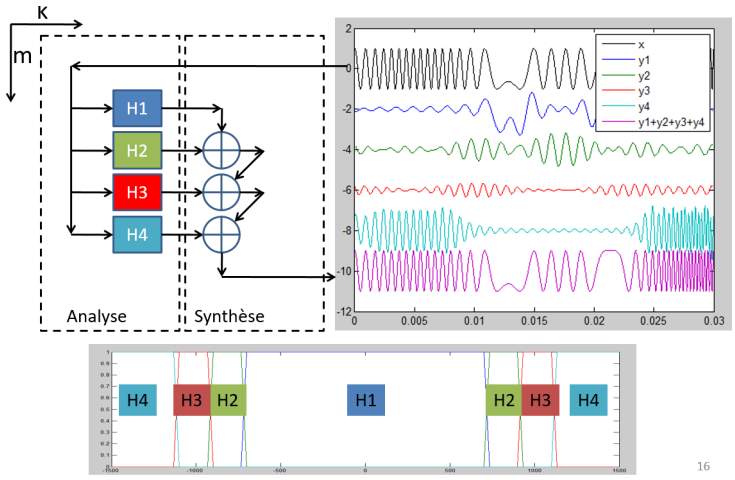

Figure 3.7: Schéma d'un banc de filtres

On introduit de nouvelles notations pour décrire un banc de filtres.

- l est un indice qui compte les raies du spectre des trames.
- f_l est la valeur de la fréquence associée à la raie l.
- m est un indice qui numérote les filtres utilisés dans le banc de filtres.

[7] Cet estimateur est défini dans la section 6.1.3 p. 13 de [?].
[8] Cet estimateur est défini dans la section 6.1.4 p. 14 de [?].

- $\widehat{H}_m(f)$ est la réponse fréquentielle du filtre numéro m.

La figure 3.7 schématise un banc de filtres. Le bas de la figure montre les réponses fréquentielles de 4 filtres notés $\widehat{H}_1(f)$, $\widehat{H}_2(f)$, $\widehat{H}_3(f)$ et $\widehat{H}_4(f)$. Le filtre $\widehat{H}_1$ est un passe-bas. Les filtres $\widehat{H}_2(f)$ et $\widehat{H}_3(f)$ sont des filtres passe-bande. Le filtre $\widehat{H}_4(f)$ est un filtre passe-haut. Une particularité de ces filtres est que la somme de leur réponses fréquentielles est un filtre passe-tout généralement un simple filtre à retard dont le module de la réponse fréquentielle vaut 1.

Dans la partie supérieure de la figure 3.7, on retrouve à gauche les 4 filtres $\widehat{H}_1(f)$, $\widehat{H}_2(f)$, $\widehat{H}_3(f)$ et $\widehat{H}_4(f)$. Ces filtres reçoivent tous en entrée le signal $x(t)$, celui-ci est visualisé en haut à droite en noir. Le signal en bleu représenté juste en dessous est $y_1(t)$, c'est la sortie du filtre de réponse fréquentielle $\widehat{H}_1(f)$. On observe que ce signal présente des variations de moindre amplitudes, c'est la conséquence de ce que $\widehat{H}_1(f)$ est un passe-bas. Les signaux représentés en dessous en vert, rouge et cyan, sont notés $y_2(t), y_3(t), y_4(t)$. Ce sont respectivement les sorties des filtres $\widehat{H}_2(f)$, $\widehat{H}_3(f)$ et $\widehat{H}_4(f)$. On y retrouve des variations parfois significativement plus importantes. La décomposition du signal $x(t)$ en 4 signaux $y_1(t), y_2(t), y_3(t), y_4(t)$ est appelé *analyse*, terme indiqué en bas à gauche de la figure 3.7. Ces signaux, $y_1(t), y_2(t), y_3(t), y_4(t)$ sont ensuite additionnés pour former le signal visualisé en bas à droite en violet et noté $y_1(t)+y_2(t)+y_3(t)+y_4(t)$. Tel qu'on peut l'observer, ce signal est identique à $x(t)$. En réalité, ce dernier signal est généralement retardé par rapport à $x(t)$. Il est en effet impossible de réaliser une décomposition en banc de filtres avec des filtres causaux sans introduire un retard. La formation d'une reconstruction retardée de $x(t)$ à partir des signaux $y_1(t), y_2(t), y_3(t), y_4(t)$ s'appelle *synthèse*, ce qui est indiqué en bas du deuxième bloc.

En regardant plus attentivement le bas de la figure 3.7, on peut observer que les supports [9] des réponses fréquentielles des différents filtres ne sont pas disjoints, on dit que ces filtres se chevauchent. En fait l'allure même du signal $y_1(t)$ avec ces variations atténuées mais existantes laisse penser qu'il y a probablement un chevauchement assez important. Et en effet la notion de banc de filtre n'est pas du tout contradictoire avec un chevauchement même assez important. Pour la suite de ce document et pour plus de simplicité, on va considérer des bancs de filtres sans chevauchement où les réponses fréquentielles sont en fait des portes définies par $\mathbf{1}_{[f_1,f_2]}(f)$.

On dit qu'un banc de filtres suit une échelle linéaire lorsque les intervalles associés aux différentes portes caractérisant les réponses fréquentielles des filtres $\widehat{H}_m(f)$ sont de longueur identiques.

$$[f_{\min}, f_{\max}] = \bigcup_m [f_m, f_{m+1}] \quad \text{où} \quad \begin{cases} f_{\min} = f_1, \ f_{\max} = f_{M+1} \\ f_m = f_{\min} + (m-1)\frac{f_{\max}-f_{\min}}{M} \\ \widehat{H}_m(f) = \mathbf{1}_{[f_m, f_{m+1}]}(f) \end{cases} \tag{3.14}$$

La sensibilité de l'oreille humaine vis-à-vis de la décomposition fréquentielle d'un son et la façon dont les sons sont produits par les organes humains amènent à choisir pour les intervalles caractérisant les filtres $\widehat{H}_m(f)$ une échelle logarithmique. Ce terme d'échelle logarithmique signifie que les intervalles sont de longueurs *croissantes* et que les bornes des ces intervalles suivent une loi *géométriques*.

$$[f_{\min}, f_{\max}] = \bigcup_m [f_m, f_{m+1}] \quad \text{où} \quad \begin{cases} f_{\min} = f_1, \ f_{\max} = f_{M+1} \\ f_m = f_{\min}\left(\frac{f_{\max}}{f_{\min}}\right)^{\frac{m-1}{M}} \\ \widehat{H}_m(f) = \mathbf{1}_{[f_m, f_{m+1}]}(f) \end{cases} \tag{3.15}$$

En pratique pour un banc de filtre suivant une échelle linéaire, on peut choisir $f_{\min} = 0$ et $f_{\max} = \frac{f_e}{2}$, et pour une échelle logarithmique, il faut que $f_{\min} > 0$, on peut choisir $f_{\min} = \frac{1}{T_K}$ et $f_{\max} = \frac{f_e}{2}$, T étant la durée de la trame.

La figure 3.8 montre les réponses fréquentielles de huit filtres $\widehat{H}_m$ qui suivent une échelle logarithmique, à gauche en les superposant sur un même graphe 2D et à droite en mettant chaque courbe le long d'un axe différent.

La figure 3.9 représente un exemple de représentation de ce descripteur temps-fréquence, il s'agit à gauche et à droite des courbes qui dans la figure 3.8 sont appelées $y_1(t), y_2(t), y_3(t), y_4(t)$. La différence est qu'ici il y en a 5 et que les filtres utilisés ici ne se chevauchent pas. On observe des variations d'amplitudes plus faibles pour les quatre premières par rapport à la dernière associées au fréquences plus élevées. Cependant il est significatif que la courbe bleue foncée, associée au filtre passe-bas et située au milieu sur le graphe de gauche et en haut à gauche sur le graphe

[9] Par support, on entend ici l'intervalle sur lequel une fonction ici la réponse fréquentielle est non nulle.

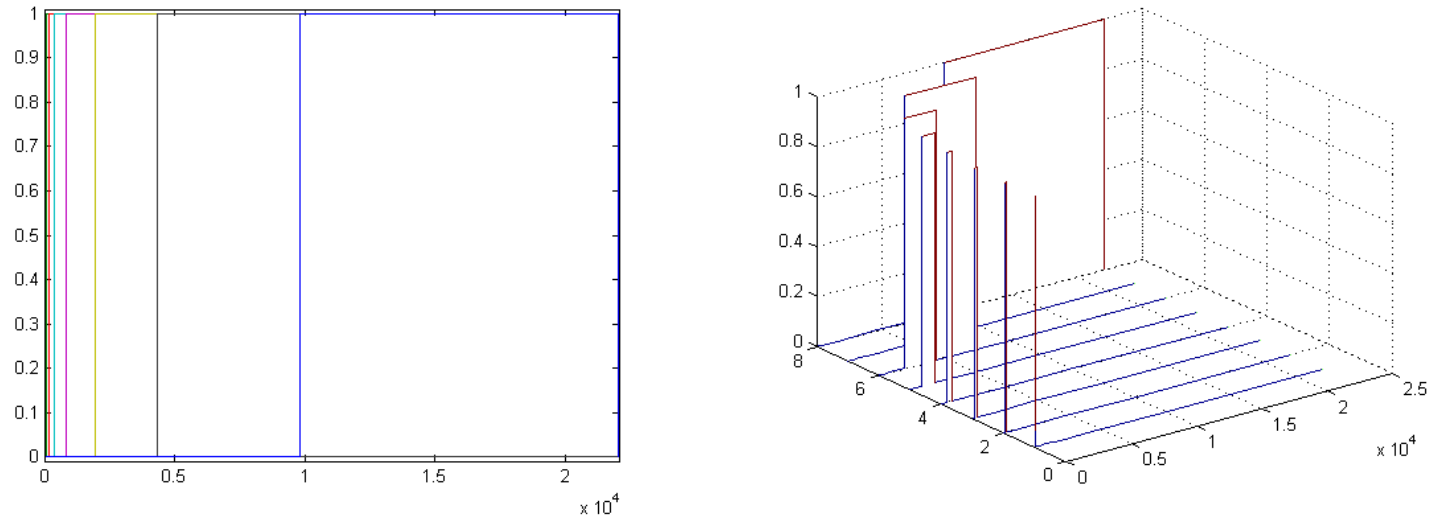

Figure 3.8: À gauche : les courbes $\widehat{H}_m(f)$ en fonction de f pour $m \in \{1, \ldots, 8\}$ sont superposées sur un même graphe. À droite : ces courbes ont chacune leur abscisse.

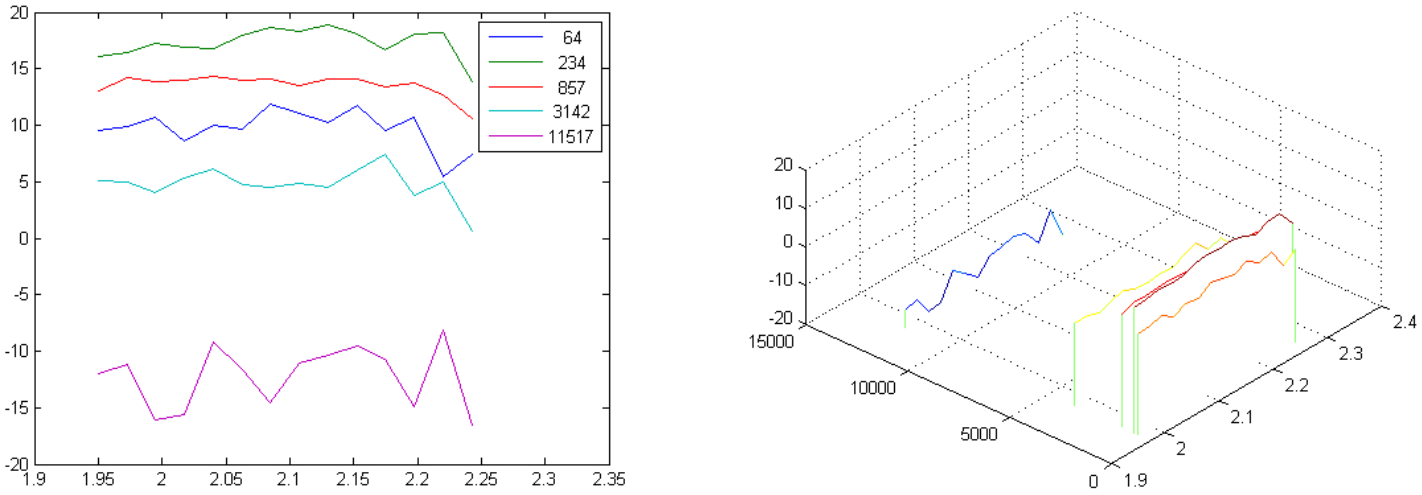

Figure 3.9: À gauche : superposition de l'évolution des descripteurs en fonction du temps. Chaque courbe est associée à un filtre en particulier dont la légende indique la fréquence centrale associée à ce filtre. À droite : représentation séparée de l'évolution de chaque descripteur en fonction du temps.Dans les deux graphes, l'échelle en temps indique les instants associés aux milieux de chaque trame.

de droite, a des variations qui sont d'une amplitude assez significatives, là on aurait pu s'attendre à des variations moins importantes. Cela traduit le fait que le signal sonore étudié présente des variations temporelles importantes d'une trame à une autre.

Ces bancs de filtres servent à définir des descripteurs, un par filtre du banc de filtre. De multiples descripteurs sont proposés dans la littérature scientifique. On aurait pu s'attendre à définir un descripteur comme la puissance sur une portion du signal en sortie du filtre $\widehat{H}_m(f)$ autour de l'instant t.

$$\boxed{D_m(t) = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} |h_m(\tau) * x_t(\tau)|^2 \, d\tau \quad \text{où} \quad \begin{cases} h_m(\tau) \text{ est la réponse impulsionnelle du filtre } \widehat{H}_m(f) \\ x_t(\tau) = x(\tau + t)\mathbf{1}_{\left[-\frac{T}{2},\frac{T}{2}\right]}(\tau) \end{cases}} \tag{3.16}$$

Mais comme la perception de l'intensité sonore est mieux corrélée à une intensité mesurée en décibel et donc suivant une échelle logarithmique vis-à-vis de la puissance, il convient d'ajouter un logarithme dans l'équation (3.16). La littérature scientifique a fait le choix d'utiliser un autre descripteur dont une version en temps continu serait

$$\text{D\_LOG}_m(t) = \frac{1}{f_{m+1} - f_m} \int_{f_m}^{f_{m+1}} 10 \log_{10} \left| \widehat{H}_m(f) \widehat{X}_t(f) \right|^2 df \tag{3.17}$$

Et en pratique, le descripteur utilisé, dépend de deux paramètres, m pour le filtre et par suite la fréquence associée à ce filtre, et k pour la trame considérée.

Le descripteur utilisé dans les travaux pratiques a pour nom `LOG_MAG_FB_LOG` [10]

En pratique, le descripteur est ainsi défini.

$$\boxed{\text{LOG\_MAG\_FB\_LOG}(m,k) = \frac{1}{|\mathbb{L}_m|} \sum_{l \in \mathbb{L}_m} 10 \log_{10} \left| \widehat{H}_{m,l} \widehat{X}_{k,l} \right|^2} \tag{3.18}$$

où $\mathbb{L}_m$ désigne les indices l pour lesquels $\widehat{H}_m\left(\frac{lf_e}{N_K}\right)$ est non-nul, ce sont a priori les indices l pour lesquels la fréquence $\frac{lf_e}{N_K}$ sont contenus dans $[f_m, f_{m+1}]$. $|\mathbb{L}_m|$ désigne le nombre d'indices l contenus dans $\mathbb{L}_m$.

Remarquons que le fait de modifier un descripteur en le multipliant par une certaine constante ou en lui ajoutant une constante n'a aucune incidence sur la classification qui en résulte.

D'autres suggestions de descripteurs pour la reconnaissance de la parole se trouvent dans [**?**].

[10] La signification du nom se retrouve en lisant l'expression de droite à gauche. `LOG` en dernier signifie qu'on considère une échelle logarithme, puis que cette échelle permet de définir un banc de filtre (*filter bank*), qu'on calcule ensuite le module (*magnitude*) et qu'avant de considérer la moyenne on considère le logarithme des modules.

Appendix A

Onde sonore stationnaire à une dimension

L'application [1] du principe fondamental de la dynamique et de la thermodynamique des fluides permettent d'écrire

$$\frac{1}{c^2}\frac{\partial^2 p(x,t)}{\partial t^2} - \frac{\partial^2 p(x,t)}{\partial x^2} = 0 \tag{A.1}$$

où c est la vitesse du son.

Les solutions de cette équation différentielle sont de la forme

$$p(x,t) = p_0 + f_1(t - \frac{x}{c}) + f_2(t + \frac{x}{c}) \tag{A.2}$$

On considère une onde avec une seule fréquence f_0. L'équation (**??**) définie page **??** est recopiée ici.

$$p(x,t) = P_0 + \Delta P_1 \cos\left[2\pi f_0\left(t - \frac{x}{c}\right) + \varphi - \psi\right] + \Delta P_2 \cos\left[2\pi f_0\left(t + \frac{x}{c}\right) + \varphi + \psi\right]$$

On considère ici un tuyau fermé de longueur L. On suppose $x = 0$ à son extrêmité gauche et $x = L$ à son extrêmité droite. Il n'y a pas de déplacement de l'air là où le tube est fermé et cela se traduit par

$$\left.\frac{\partial p}{\partial x}\right|_{x=0} = 0 \quad \text{et} \quad \left.\frac{\partial p}{\partial x}\right|_{x=L} = 0 \tag{A.3}$$

Le calcul de ces expressions à partir de $p(x,t)$ aboutit à

$$\left.\frac{\partial p}{\partial x}\right|_{x=0} = \frac{2\pi f_0}{c}\Delta P_1 \sin\left[2\pi f_0\left(t - \frac{0}{c}\right) + \varphi - \psi\right] - \frac{2\pi f_0}{c}\Delta P_2 \sin\left[2\pi f_0\left(t + \frac{0}{c}\right) + \varphi + \psi\right] \tag{A.4}$$

Le fait que cette expression est nulle montre que $\Delta P_1 = \Delta P_2$ qui est noté ΔP et que $\psi = 0$.

$$\left.\frac{\partial p}{\partial x}\right|_{x=L} = \frac{2\pi f_0}{c}\Delta P \sin\left[2\pi f_0\left(t - \frac{L}{c}\right) + \varphi\right] - \frac{2\pi f_0}{c}\Delta P \sin\left[2\pi f_0\left(t + \frac{L}{c}\right) + \varphi\right] \tag{A.5}$$

Une relation trigonométrique permet d'écrire

$$\cos(a+b) - \cos(a-b) = -2\sin(a)\sin(b) \tag{A.6}$$

Une relation trigonométrique similaire montre que

$$\sin(a+b) - \sin(a-b) = 2\sin(b)\cos(a) \tag{A.7}$$

Aussi

$$\left.\frac{\partial p}{\partial x}\right|_{x=L} = -\frac{4\pi f_0}{c}\Delta P \cos\left[2\pi f_0 t + \varphi\right] \sin\left(\frac{2\pi f_0 L}{c}\right) \tag{A.8}$$

Cette expression est bien nulle si et seulement si

$$\sin\left(\frac{2\pi f_0 L}{c}\right) = 0 \tag{A.9}$$

Ce qui est vrai lorsque

$$f_0 = k\frac{c}{L} \tag{A.10}$$

Plus la longueur augmente plus la fréquence de base diminue.

[1] Cette annexe est inspirée de [**?**]