Signaux et Systèmes Électroniques

Analyse de la parole

Auteurs :
M. Samuel Riedo
M. Pascal Roulin

Encadrant :
M. Daniel Oberson

1 Introduction

Le son de notre voix se fait grâce à nos poumons, nos cordes vocales et notre caisse de résonance, composée de notre cavité nasale et buccale. Cette dernière est propre à chaque individu et altère le timbre de la voix d'une personne alors que les muscles du larynx permettent de modifier la hauteur d'un son. Grâce à un microphone relié à un oscilloscope, il nous est possible d'analyser différentes caractéristiques de la voix ou d'un phonème, comme le ferait un système de reconnaissance vocale.

2 Analyse

2.1 Propriétés de la parole

Tout signal peut être analysé soit dans l'espace temporel, soit dans l'espace fréquentiel. Dans le premier cas, l'amplitude du signal est étudiée en fonction du temps.
Définir un signal dans l'espace fréquentiel consiste à trouver son spectre, calculé au moyen de la transformation de Fourier. Le spectre d'un signal représente toutes les sinusoïdes, qui, additionnées, constituent ledit signal. Lorsque ces fréquences sont des multiples d'une même fréquence, appelée fréquence fondamentale, les autres sont des harmoniques. De ce fait, seuls les signaux périodiques ont une fréquence fondamentale.

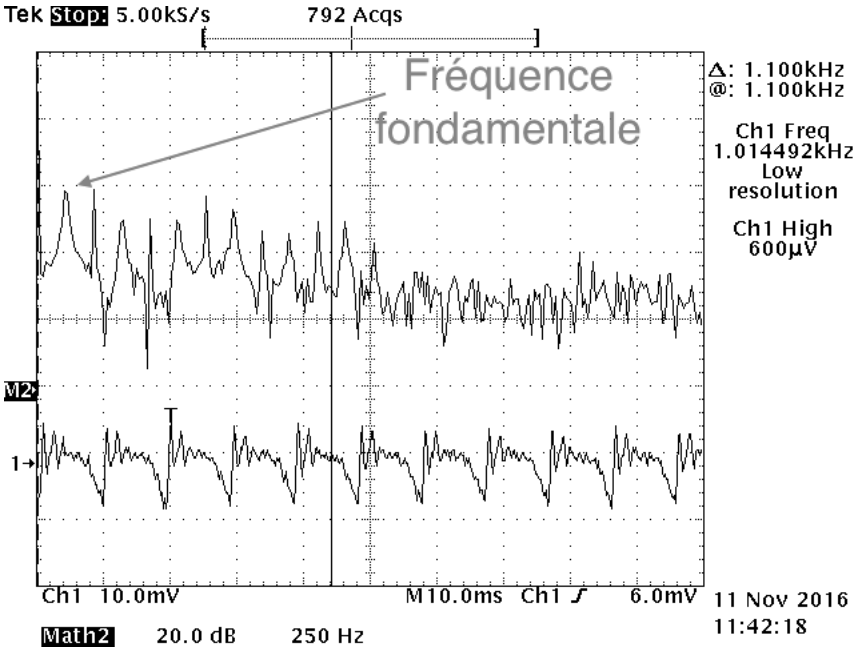


FIGURE 1 – Représentation d'un son et de son spectre

La figure 1 représente le signal du son "a" répété en boucle, le signal supérieur étant son spectre. Nous pouvons constater que ce signal est périodique avec $t \approx 10\text{ms}$. Il est donc possible de calculer la fréquence fondamentale de ce signal :

$$\begin{aligned} F_{fondamentale} &= \frac{1}{10\text{ms}} \\ &= 100\text{Hz} \end{aligned}$$

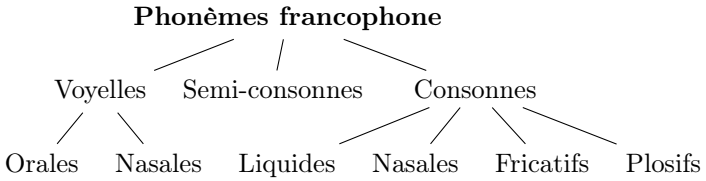
Si nous analysons le spectre du signal, la fréquence fondamentale est censée être la première harmonique, les autres étant des multiples de cette dernière. L'axe y correspond à la fréquence, par pas de 250Hz. Le premier pic du spectre se trouve à environ 0.4 div, nous pouvons donc lire que la fréquence fondamentale correspond à :

$$\begin{aligned} F_{fondamentale} &= 0.4 \cdot 250\text{Hz} \\ &= 100\text{Hz} \end{aligned}$$

Ce qui donne le même résultat que lors du premier calcul.

2.2 Phonèmes

Un phonème est la plus petite unité distinctive isolable dans le signal d'un mot. En d'autres termes, chaque mot est une suite de phonème. Il est possible de les classer selon leur type.



En plus de toutes ces caractéristiques, chaque phonème est **voisé** ou **non voisé**. La première différence entre les deux est simplement qu'un phonème voisé fera vibrer les cordes vocales. D'un point de vue traitement du signal, les phonèmes voisés peuvent être différencié selon leur formant (voir point 2.3).

— **Plosif :** Son court, mais intense. La bouche est fermée au début, puis s'ouvre d'un coup, comme si elle "explosait". Par exemple, les sons *b, p* et *t* sont plosifs. Il est humainement impossible de répéter un phonème plosif en continu sans faire de pause entre-deux, ou de dire un autre phonème. Par exemple, pour répéter le phonème *B* en boucle, une personne aura tendance à dire *Bééé*, ce qui correspond à un deux phonèmes (*B* et *é*).

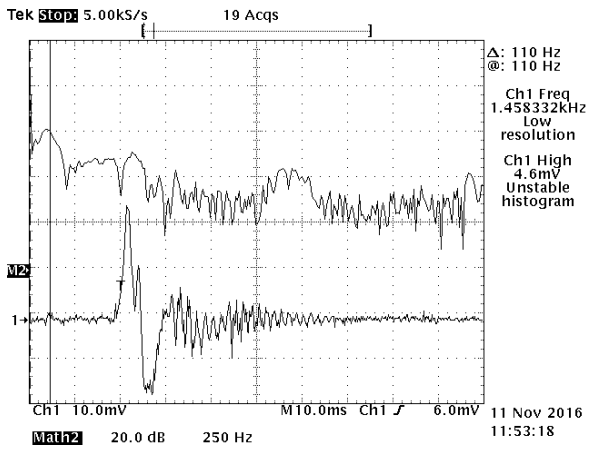


FIGURE 2 – Phonème plosif "B" voisé

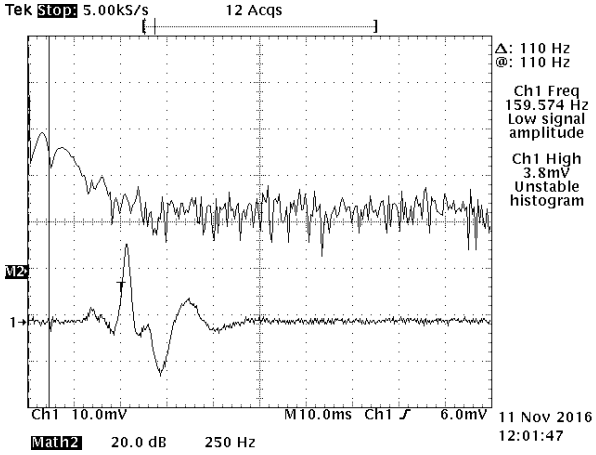


FIGURE 3 – Phonème plosif "P" non-voisé

Il est intéressant de noter que le spectre du phonème voisé à tendance à ne pas rester régulier en amplitude après l'impulsion du signal, au contraire du spectre du phonème *P* non voisé. Cette différence importante est liée aux vibrations des sorbes vocales absentes dans un cas non voisé, et aura une grande importance lorsqu'il s'agira de traiter les signaux afin de reconnaître le phonème lié à ces signaux.

— **Fricatif :** Il s'agit d'un soufflement. Aucune partie du corps ne résonne ou ne vibre. La lettre *s* est le phonème le plus fricatif, il contient toutes les fréquences. Pour cette raison, ces phonèmes sont également souvent appelés bruit ou noise.

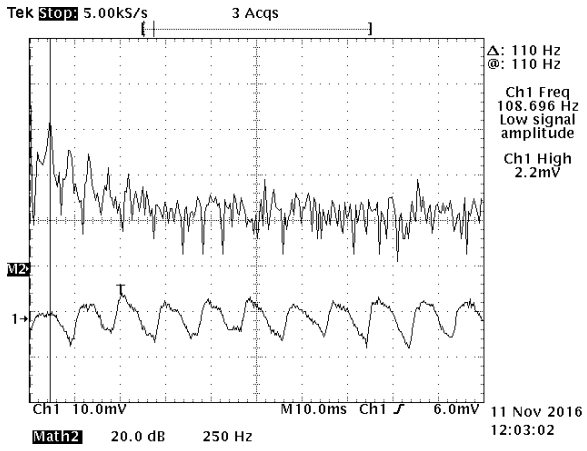


FIGURE 4 – Phonème fricatif "V" voisé

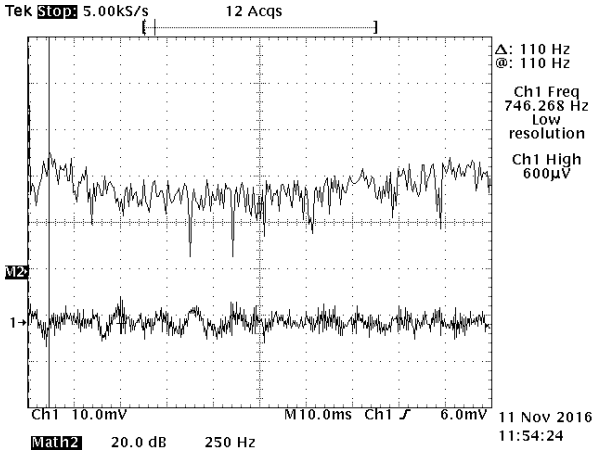


FIGURE 5 – Phonème fricatif "CH" non voisé

— **Nasale :** Résonance dans le nez, par exemple avec la lettre *m*.
— **Liquide :** La langue influence le son produit.

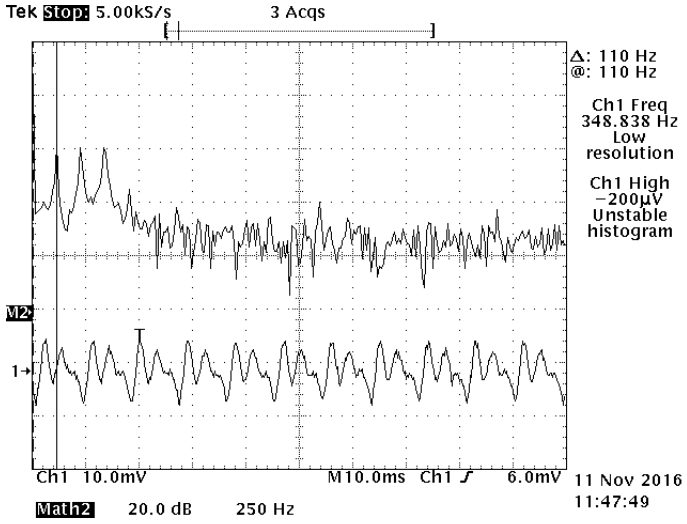


FIGURE 6 – Lettre "l" répétée en boucle

— **Orales :** À l'opposé d'un son nasal, la caisse de résonance ici est la bouche. Les lettres *i, e* et *a* sont par exemple des phonèmes oraux.

2.3 Formant

Durant les dernières années sont apparus plusieurs outils de reconnaissance vocale tels que Siri par Apple ou Cortana par Microsoft. Ces programmes enregistrent au moyen d'un micro une question posée oralement par un humain afin de lui donner une réponse. La principale difficulté derrière ces programmes n'est pas de répondre à la demande de l'utilisateur, mais de transformer le signal audio en un texte au format binaire qu'une machine peut interpréter. Ainsi, comment différencier le signal d'un *a* de celui d'un *o* ?
Nous avons vu précédemment que tout mot est une suite de phonèmes, et que ses derniers sont soit voisés, soit non voisés. Pour la première catégorie, il est possible de les différencier selon leur formant. Par définition, un formant est un concentré d'énergie acoustique autour d'une fréquence particulière d'une onde. Chaque formant correspond à une résonance, ce qui explique pourquoi seuls les phonèmes voisés en ont.

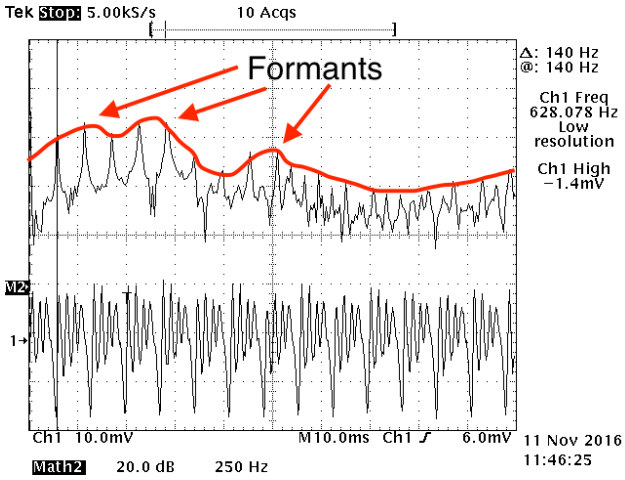


FIGURE 7 – Formants de la lettre A

Sur le spectre, les principales concentrations d'énergie se trouvent à 625Hz et 1250Hz. Le premier formant est volontairement ignoré, car il représente le pitch. Habituellement, une voyelle a trois formants, mais le signal ci-dessus est trop court pour voir le dernier. Le tableau ci-dessous (figure 8) permet de trouver le phonème correspondant au spectre en utilisant la fréquence des deux premiers formants.

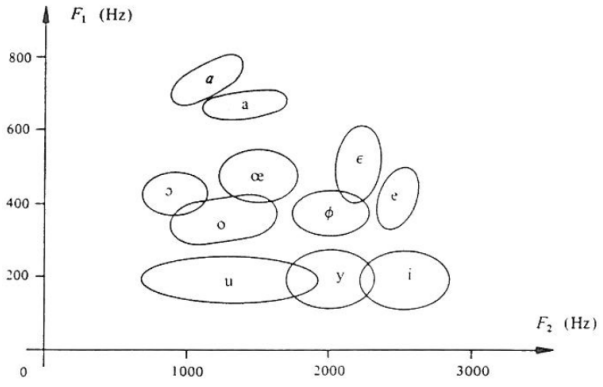


FIGURE 8 – Représentation des voyelles sur le plan fréquentielle

En plaçant 625Hz sur l'axe f_1 et 1250Hz sur l'axe f_2, le croisement des valeurs tombe sur l'ellipse correspondant au phonème *a*, ce qui correspond bien à ce qui a été prononcé dans le micro. Les formants permettent donc de distinguer tous les phonèmes voisés entre eux.

2.4 Pitch

Chaque être humain à une voix différente et unique utilisable comme un identifiant, à la manière des empreintes digitales. Elle est composée de plusieurs composantes, telles que le pitch ou le ton.
Le pitch de la voix est défini comme étant la fréquence à laquelle vibrent les cordes vocales. Le son produit change selon cette même fréquence. Généralement, les enfants ont une voix plus aigüe que les adultes. Cela vient du fait que leurs cordes vocales vibrent à une fréquence plus élevée.
La fréquence de vibration des cordes vocales dépend principalement de leur longueur et de leur épaisseur, mais aussi de l'état des muscles les entourant. C'est sur ce dernier paramètre qu'un humain peut influencer pour modifier sa voix et imiter quelqu'un. De même, lorsque l'on est heureux ou triste, la voix change du fait qu'inconsciemment nos muscles se contractent et se relaxent.
Nous avons vu au point 2.1 que la fréquence fondamentale d'un signal était la plus basse, et que toutes ses harmoniques étaient des multiples de cette dernière. Comme les sons produits par les cordes vocales dépendent de la fréquence à laquelle vibrent les cordes vocales, nous pouvons en déduire que le pitch est égal à la fréquence fondamentale dans ce cas.

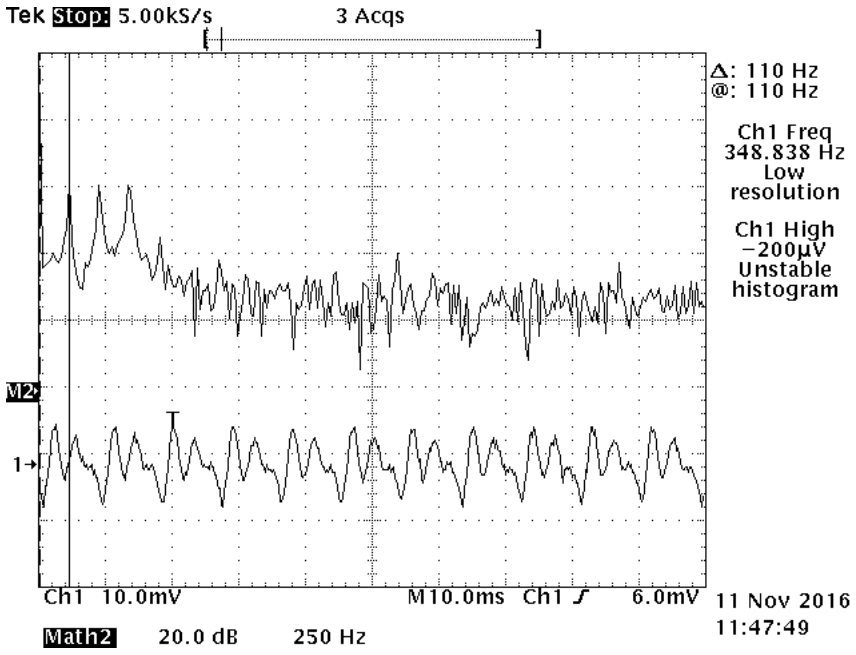


Figure 9 – Lettre "l" répétée en boucle

Sur la figure 9, un individu prononce la lettre *l* de façon continue, le signal inférieur est donc périodique. Sur le spectre, un curseur a été placé sur le premier pic afin de lire le pitch (le pic tout à gauche de l'image est la composante continue du signal). Le pitch de cette personne vaut donc 110Hz.

Nous avons vu que ce pitch dépendait de la longueur des cordes vocales, de leur épaisseur ainsi que de l'état des muscles les entourant. Si la même personne prononce maintenant la lettre *o* en série, le pitch ne serait donc pas sensé changer sauf si elle était d'humeur différente ou qu'elle se forçait.

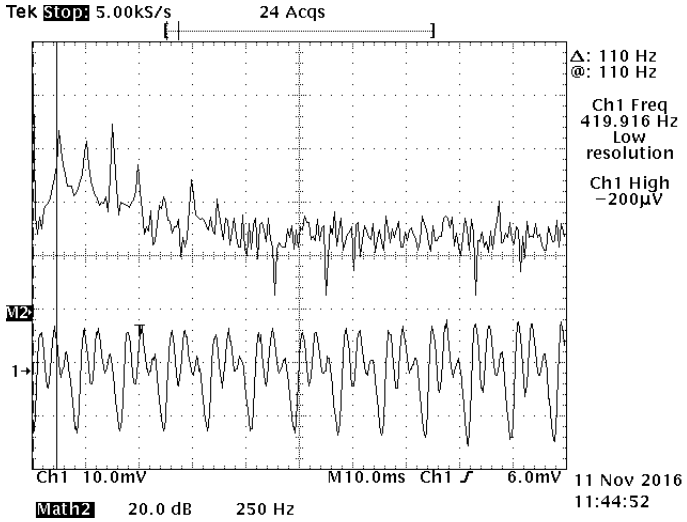


FIGURE 10 – Lettre "o" répétée en boucle par la même personne

Le pitch reste en effet à 110Hz, il ne dépend donc pas du phonème. Si une personne différente prononçait le phonème *a*, il y a de grandes chances qu'il soit différent de 110Hz.

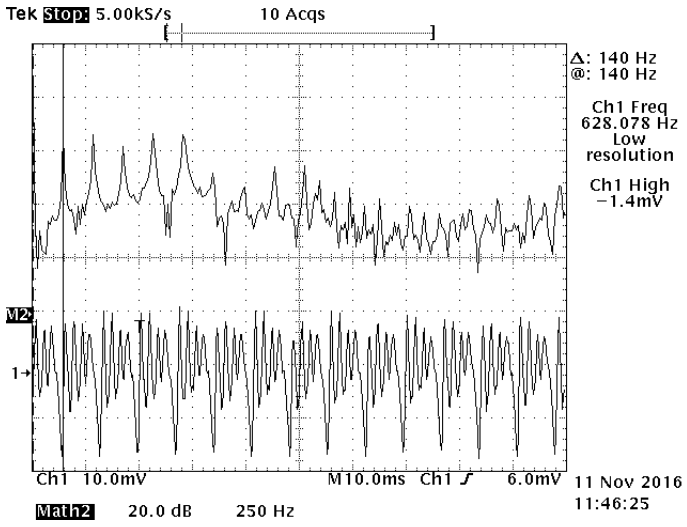


FIGURE 11 – Lettre "a" répétée en boucle par une personne différente

Cette fois, le pitch vaut 140Hz. Nous pouvons donc affirmer qu'il ne s'agit pas de la même personne.

3 Conclusion

L'analyse du son est un sujet très complexe et important. Aujourd'hui, des ressources sont massivement investies dans le développement d'intelligences artificielles utilisant la reconnaissance vocale. La reconnaissance vocale passe par de nombreuses étapes : découpage des phrases en mots, puis des mots en phonèmes, la reconnaissance de ces derniers variant selon leur type.
Un oscilloscope et un analyseur de spectre nous permettent de mieux comprendre de quoi les sons sont constitués, et comment les différencier. Ces dernières années, des avancées majeures ont été effectuées, mais il est encore trop tôt pour réussir à avoir une reconnaissance vocale fonctionnant sur un terminal tel qu'un téléphone sans sous-traiter le travail à un serveur distant.

Samuel Riedo

Pascal Roulin