

# Réflexion sur la difficulté de présenter une expertise scientifique au tribunal : l'exemple de l'analyse de la voix au Québec

Yvan Dutil, PhD , phys.

# Une demande particulière

- Je suis physicien et j'ai fait (en autre) de la consultation légale en physique.
- Un jour un avocat me demande d'analyser un fichier audio
  - Son client nie avoir eu une conversation qui l'incrimine.
  - Le témoin qui a fourni l'enregistrement s'est déjà parjuré dans le passé.
  - Un ingénieur du son a regardé le fichier et n'a rien trouvé.
- N'ayant jamais fait cela, j'essaye de trouver une autre personne pour le faire
  - N'ayant trouvé personne et ayant examiné la littérature, et m'étant convaincu que je pouvais faire un travail correct, j'accepte le contrat.

# Premières difficultés

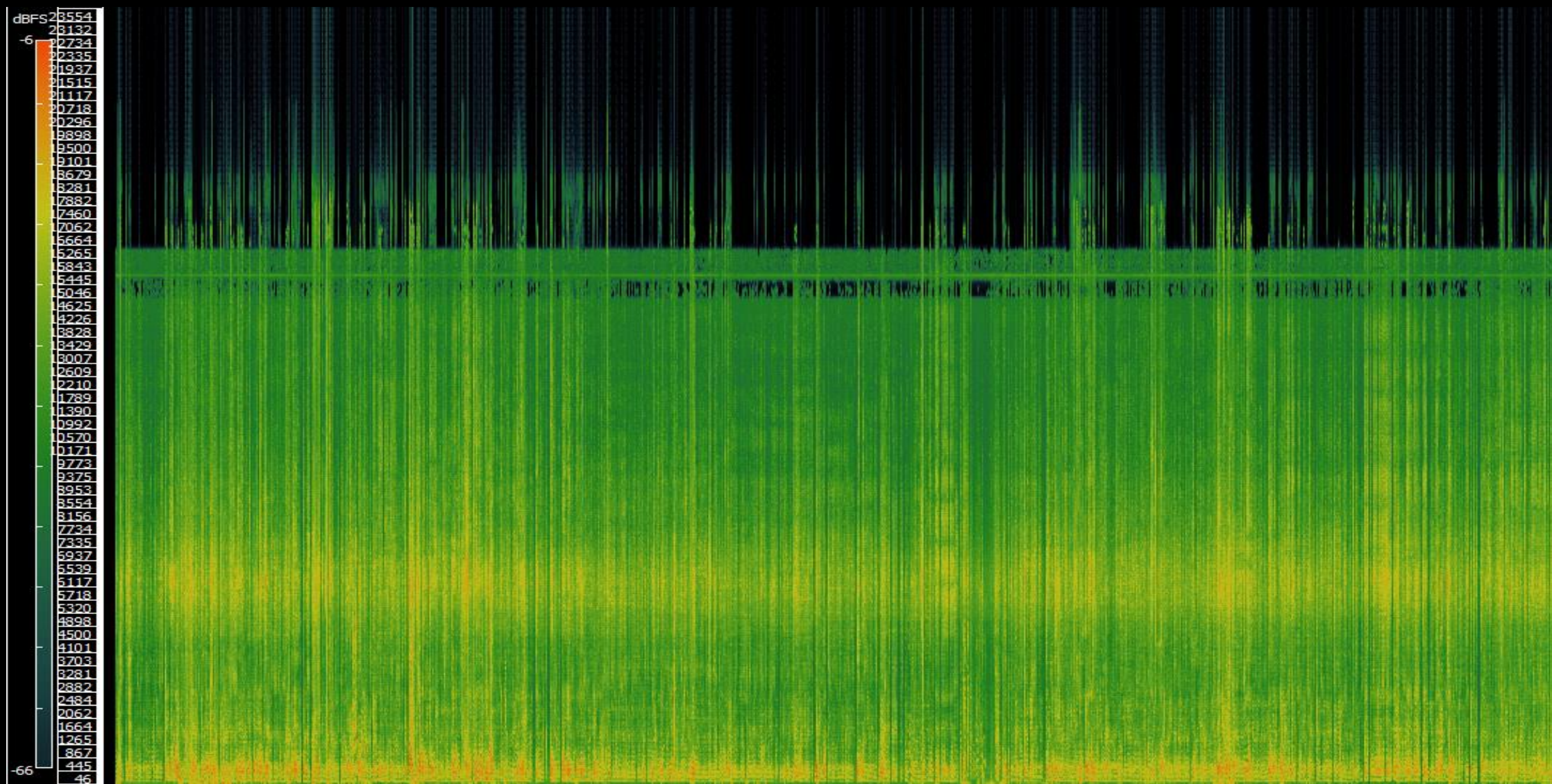
- Les professionnels ont accès à des outils informatiques spécialisés très coûteux.
- Je dois contourner la difficulté
  - Utiliser des logiciels de domaines publics servant à l'analyse de la voix
  - Écrire mes propres algorithmes d'analyse à partir de mes connaissances en traitement de signal.
- Économique, mais très coûteux en temps

# Premières difficultés

- Étant donné les doutes sur la véracité du contenu, il faut faire un maximum de vérification avant même de commencer l'analyse.
  - Est-ce que l'histoire se tient?
  - Est-ce que l'enregistrement est consistant physiquement?
- L'enregistrement a été fait à partir d'un téléphone cellulaire.
  - Plusieurs années se sont écoulées.
  - Le téléphone n'existe plus.
  - Le témoin qui a enregistré la conversation ne se souvient pas du modèle (Galaxy ...)!
  - Impossible de retourner sur les lieux de l'enregistrement!

# Description du fichier audio

- Fichier audio monopiste .wav : durée 31 m 47,247 s.
  - Converti d'un format natif selon toutes évidences.
- Taux d'échantillonnage de 48 000 Hz.
- Le signal audio n'apparaît qu'après 0,149 s.
- Même si la fréquence de Nyquist est de 24 000 Hz, il n'y a pratiquement aucune énergie à des fréquences supérieures à 15 000 Hz.
  - Suréchantillonnage par rapport au format natif.



Spectrogramme de l'ensemble du fichier audio

# Métadonnées

- Oups! On n'a pas le fichier original!
  - Pas l'heure de l'enregistrement
  - Pas la durée
  - Pas la position du téléphone
  - Pas le type du téléphone
- Heureusement, la copie n'a pas été altérée après l'extraction:  
`<xmpMM:DocumentID>`  
`xmp.did:98D761D85F4BE418AE0DBEF206B442C<`  
`/xmpMM:DocumentID>.`

# Examen préliminaire du fichier audio

- La conversation se passe entre un homme et une femme. Il y a un enfant présent dans les environs. On entend en arrière-plan ce qui semble être le son d'une émission pour enfant. À la fin de l'enregistrement, un second homme entre dans la pièce. On entend aussi des bruits extérieurs.
- Le niveau de bruit ambiant est relativement important.
  - On observe la présence de nombreux clics et coups.
  - Le bruit de fond est semblable à un bruit de machine rotative. **On dirait une laveuse (vaisselle ou linge).**
- De plus, **le bruit de fond change de façon considérable pendant la durée de l'enregistrement.**
- Les interlocuteurs parlent librement et **ne semblent pas se préoccuper de l'enregistrement.** À plusieurs reprises, on entend de nombreux clics, ce qui donne l'impression que l'enregistreur est manipulé. On entend clairement leurs respirations, ce qui indique que l'enregistreur est placé très près d'eux.

# Examen préliminaire du fichier audio

- Le fichier audio a été examiné en détail à l'aide du logiciel *Sonic Visualiser*.
- L'examen portait à la fois sur un examen de l'onde sonore et du spectre en fréquence.
- Cet examen a aussi permis de segmenter le message selon les locuteurs, isoler les silences, repérer d'éventuelles anomalies et de préparer un verbatim partiel de la conversation ainsi qu'une description des bruits environnants.

# Examen préliminaire du fichier audio

| Segmentation du fichier audio |          |             |
|-------------------------------|----------|-------------|
| Catégorie                     | Segments | Durée       |
| Locuteur                      | 372      | 15 min 54 s |
| Silence                       | 424      | 7 min 44 s  |
| Femme                         | 211      | 6 min 42 s  |
| Enfant                        | 11       | 19 s        |
| Homme                         | 3        | 5,9 s       |
| Voix superposées              | 6        | 6,6 s       |

# Notes sur l'analyse du fichier audio

|         | Yvan Dutil, PhD |          |          |       |           |                                               |
|---------|-----------------|----------|----------|-------|-----------|-----------------------------------------------|
|         |                 |          |          |       |           |                                               |
| Segment | Temps           |          |          | Durée | Catégorie | Verbatim (partiel)                            |
|         | Minutes         | Secondes | Secondes |       |           |                                               |
|         |                 |          |          |       |           |                                               |
| 1       | 0               | 0        | 0        | 1,109 | Locuteur  | Loin d'ici                                    |
| 2       | 0               | 1,109    | 1,109    | 0,981 | Silence   | Un autre spa                                  |
| 3       | 0               | 2,09     | 2,09     | 0,539 | Femme     | Pour allez où?                                |
| 4       | 0               | 2,629    | 2,629    | 1,391 | Silence   | Meilleur ou l'oreille [mot inintelligible]    |
| 5       | 0               | 4,02     | 4,02     | 0,708 | Locuteur  | y'a mom...                                    |
| 6       | 0               | 4,728    | 4,728    | 1,245 | Silence   |                                               |
| 7       | 0               | 5,973    | 5,973    | 0,784 | Locuteur  | où ailleurs                                   |
| 8       | 0               | 6,757    | 6,757    | 0,375 | Silence   |                                               |
| 9       | 0               | 7,132    | 7,132    | 0,174 | Femme     | Han?                                          |
| 10      | 0               | 7,306    | 7,306    | 0,539 | Locuteur  | Chechessss                                    |
| 11      | 0               | 7,845    | 7,845    | 0,603 | Silence   |                                               |
| 12      | 0               | 8,448    | 8,448    | 1,984 | Locuteur  | Chercher ça me rappelle trop de souvenir si.. |
| 13      | 0               | 10,432   | 10,432   | 3,191 | Silence   |                                               |
| 14      | 0               | 13,623   | 13,623   | 1,95  | Femme     |                                               |
| 15      | 0               | 15,573   | 15,573   | 0,789 | Locuteur  |                                               |
| 16      | 0               | 16,362   | 16,362   | 1,581 | Silence   |                                               |
| 17      | 0               | 17,943   | 17,943   | 0,275 | Enfant    |                                               |
| 18      | 0               | 18,218   | 18,218   | 0,619 | Silence   |                                               |
| 19      | 0               | 18,837   | 18,837   | 0,96  | Femme     | Toux femme (2x)                               |
| 20      | 0               | 19,797   | 19,797   | 1,707 | Locuteur  | (aspiration à la fin peut-être de la femme)   |
| 21      | 0               | 21,504   | 21,504   | 0,426 | Silence   |                                               |
| 22      | 0               | 21,93    | 21,93    | 1,11  | Femme     |                                               |
| 23      | 0               | 23,04    | 23,04    | 0,652 | Silence   | Voix d'enfant + plus bruit d'aspiration       |
| 24      | 0               | 23,692   | 23,692   | 1,129 | Femme     | moman?                                        |
| 25      | 0               | 24,821   | 24,821   | 0,701 | Silence   |                                               |
| 26      | 0               | 25,522   | 25,522   | 5,356 | Locuteur  |                                               |

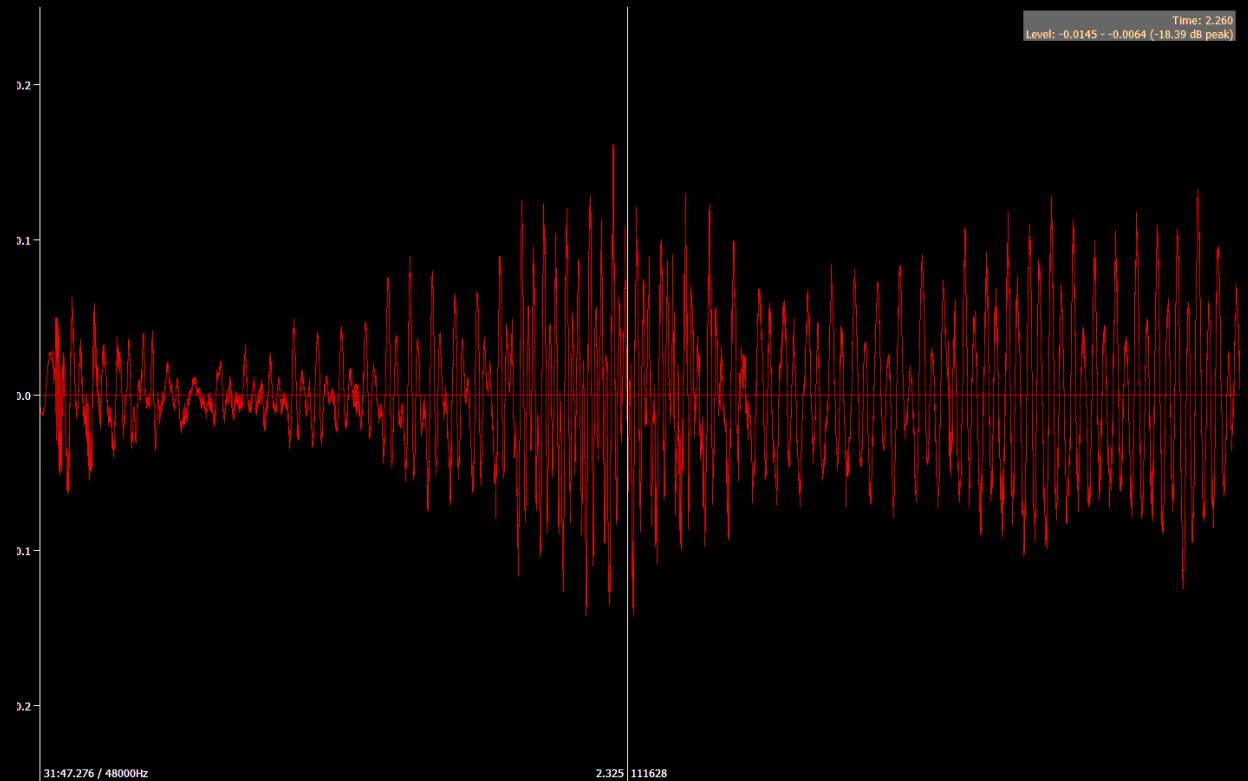
# À la recherche d'anomalie

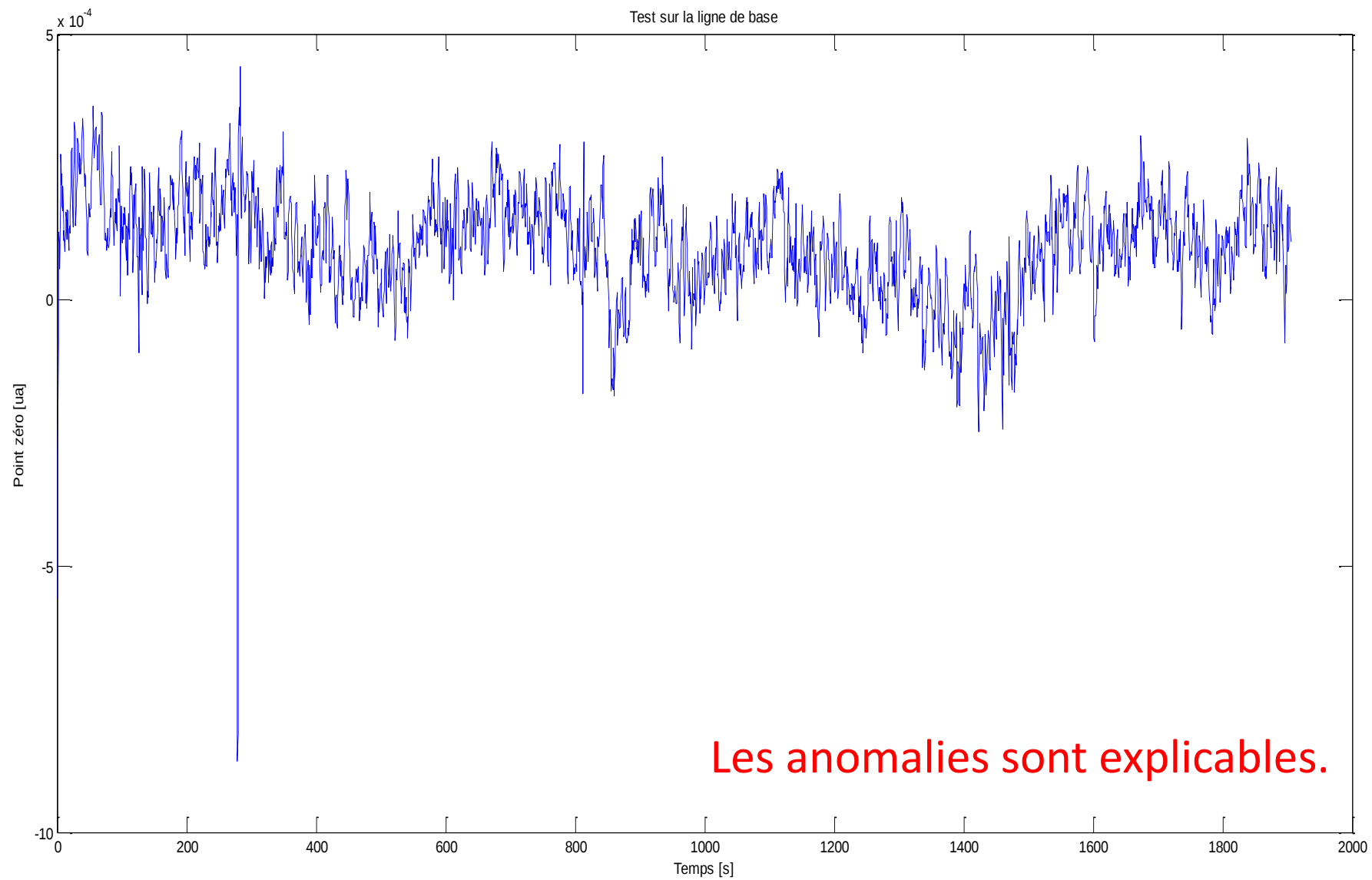
- Mon client niait avoir eu la conversation, malgré que la voix entendue ressemblait fortement à la sienne.
- Impératif de vérifier s'il n'y avait pas d'indice de la manipulation du fichier.
  - Aujourd'hui, il s'agit d'une opération relativement facile
    - **Logiciels puissants gratuits disponibles.**
- L'examen audio préliminaire ayant soulevé un doute, il fallait creuser plus loin.

# À la recherche d'anomalie

## Examen de la ligne de base.

La ligne de base reste normalement très près de zéro, mais des manipulations peuvent la déplacer.





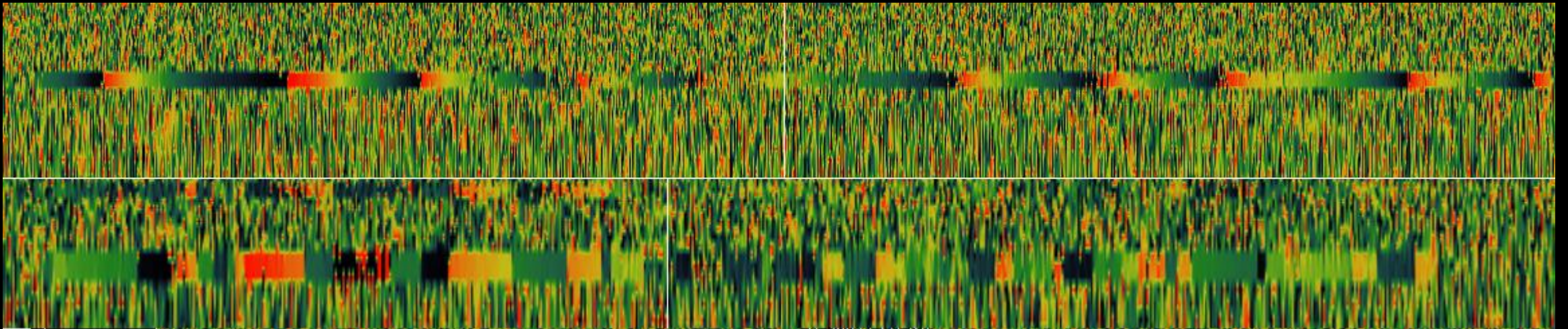
Les anomalies sont explicables.

# Étude du signal du réseau électrique

- Le signal du réseau électrique est **omniprésent**.
  - 60 Hz + harmoniques
  - Facilement détecté par les micros ou directement par l'électronique.
  - Détectable aussi sur les vidéos.
- Permet de **détecter facilement** les manipulations de fichiers

# Étude du signal du réseau électrique

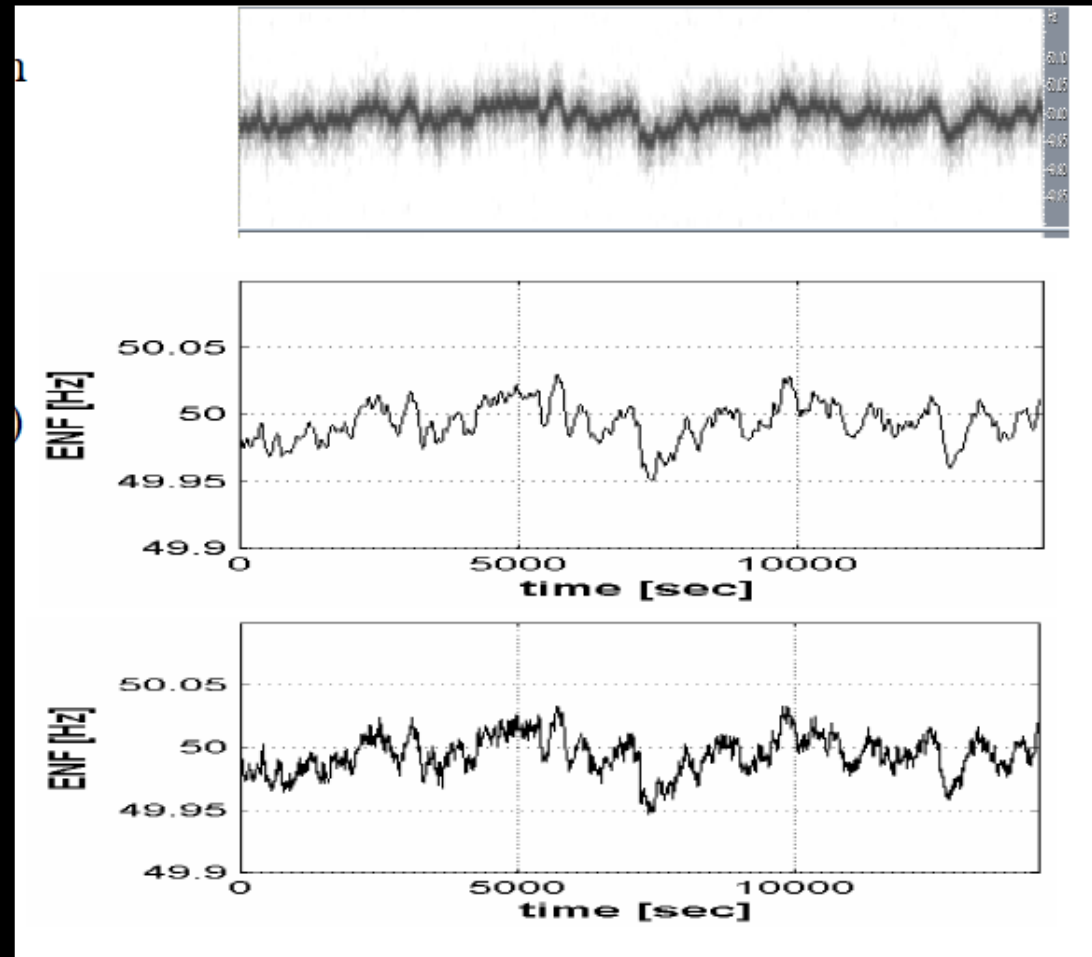
Les sauts de phases permettent de détecter facilement les manipulations



# Étude du signal du réseau électrique

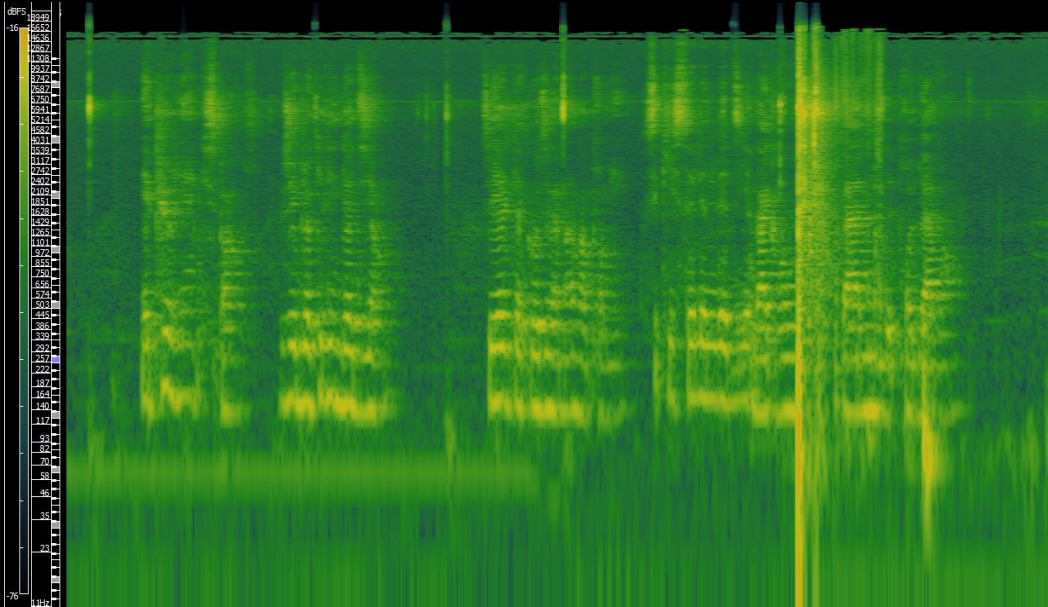
- Le signal du réseau électrique n'est pas parfaitement constant.
- Les fluctuations permettent :
  - De confirmer l'heure d'un enregistrement
  - De déterminer l'intégrité d'un enregistrement
  - Déterminer approximativement le lieu d'un enregistrement

Les données d'Hydro-Québec sont confidentielles.

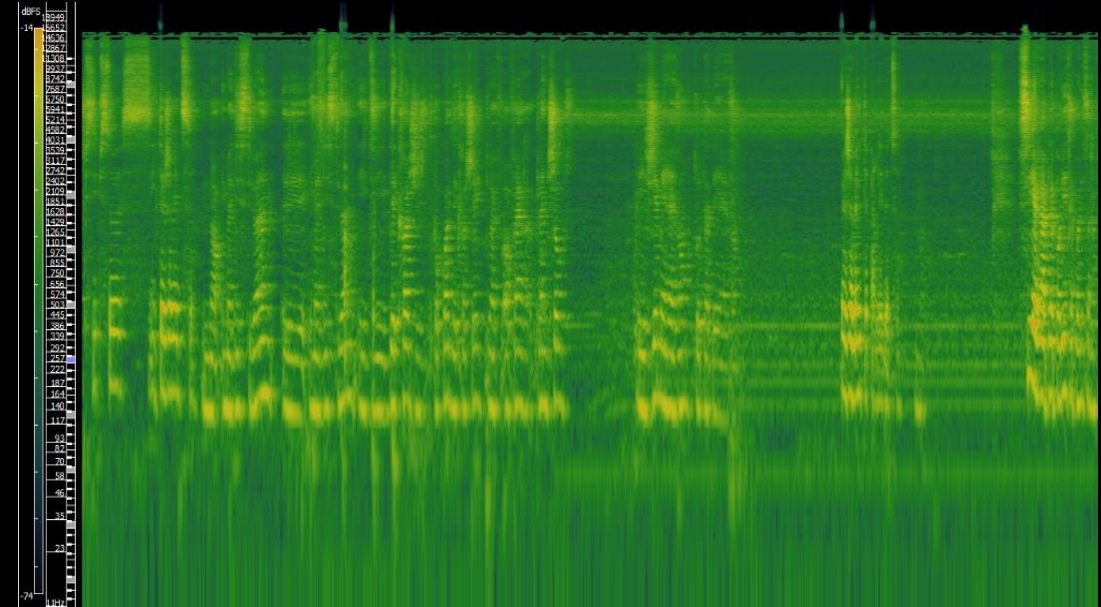


Adapté de Grigoras Catalin

# Étude du signal du réseau électrique

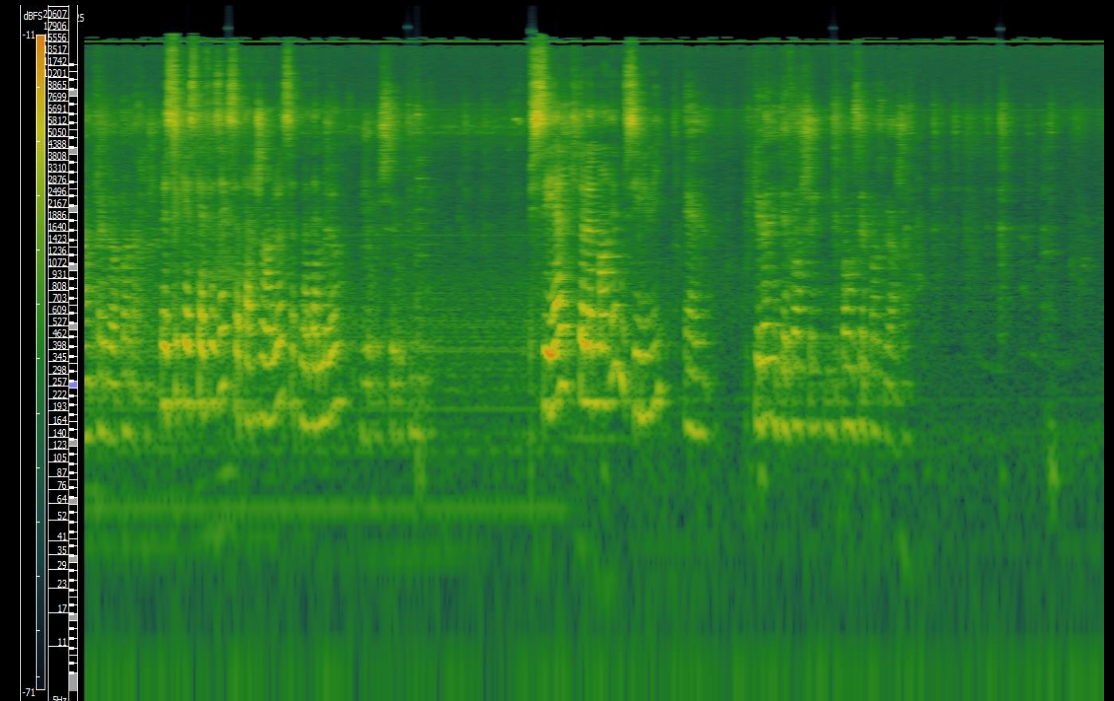
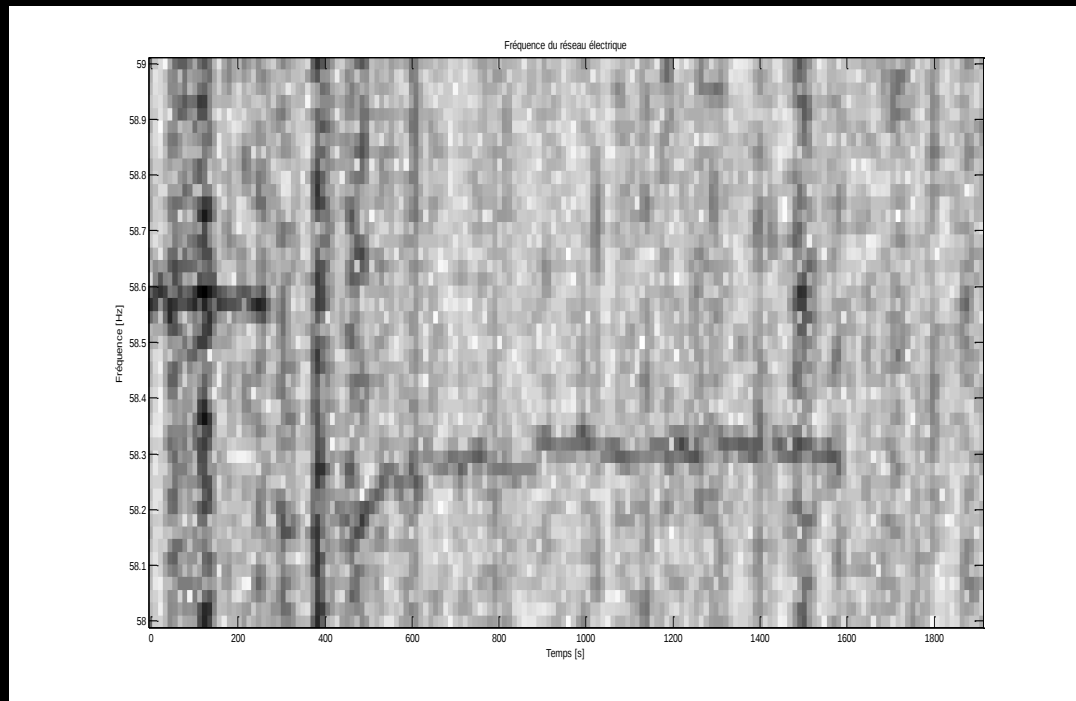


Bande de 60 Hz (4:37.23)



Bande de 60 Hz (6:13.74)

# Étude du signal du réseau électrique



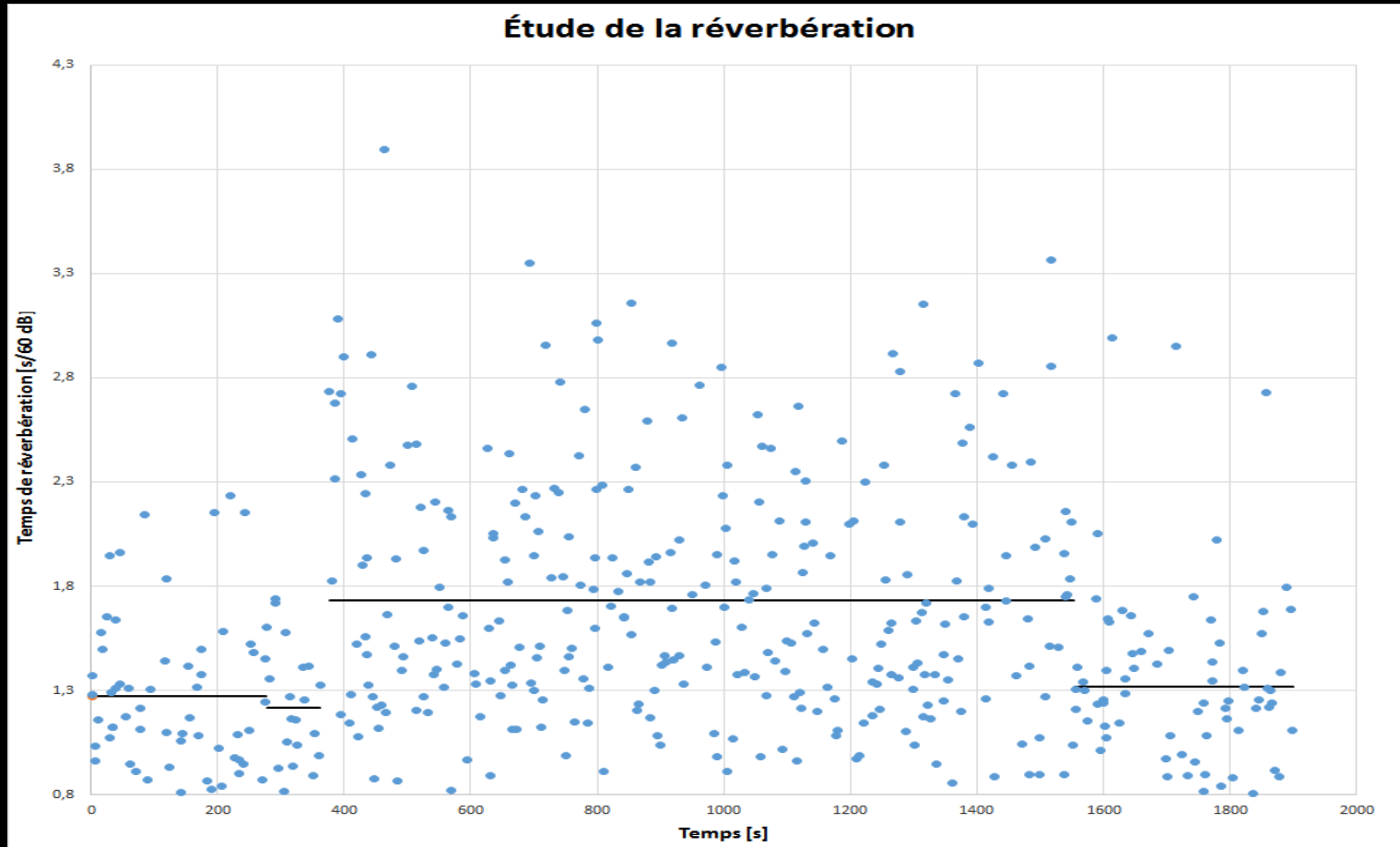
Bande de 60 Hz (25:54.47)

**Le signal électrique et le bruit de fond indiquent 4 périodes distinctes**

# Étude de la réverbération

- La réverbération du son sur les murs fait que le son ne s'éteint pas instantanément.
  - Dépend de la taille et de la géométrie de la pièce
  - Dépend du revêtement des murs.
- Empreinte physique du lieu

# Étude de la réverbération



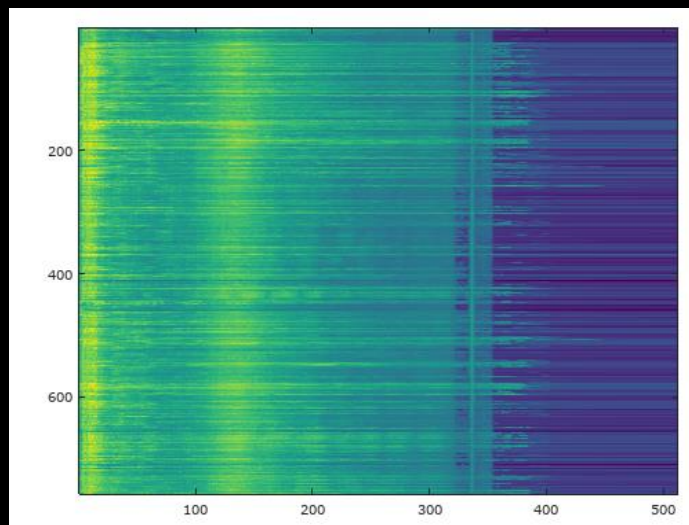
Variation  
importante  
potentiellement  
due au niveau de  
bruit interférant  
avec l'algorithme.

# Étude du bruit

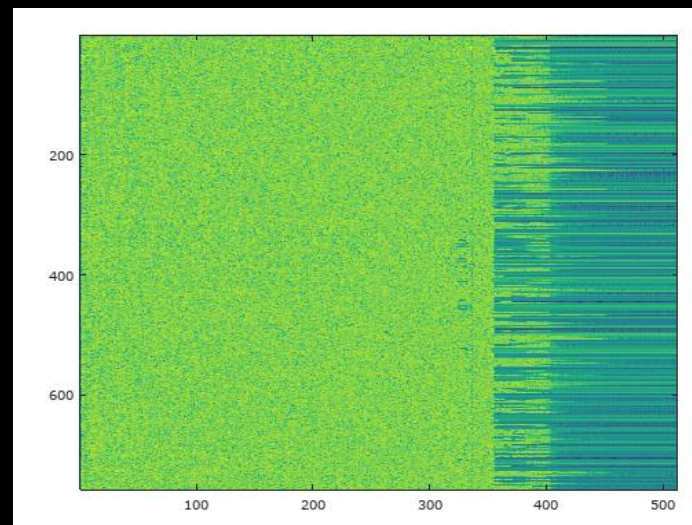
- Il est très difficile de manipuler le signal sans affecter simultanément la structure du bruit.
  - Rapport signal à bruit
  - Variance, asymétrie, aplatissement
- Pas de détection d'anomalie inexplicable.

# Analyse du bruit

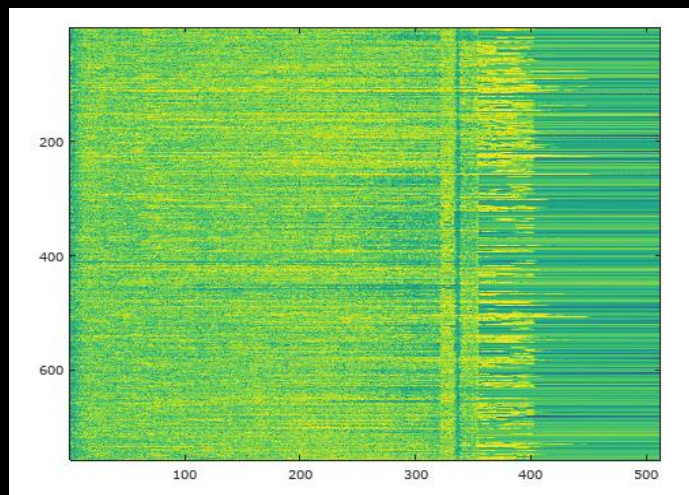
Variance



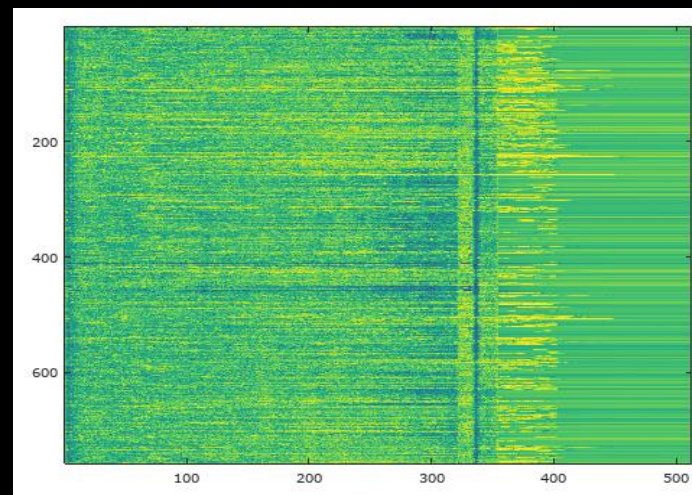
Signal/bruit



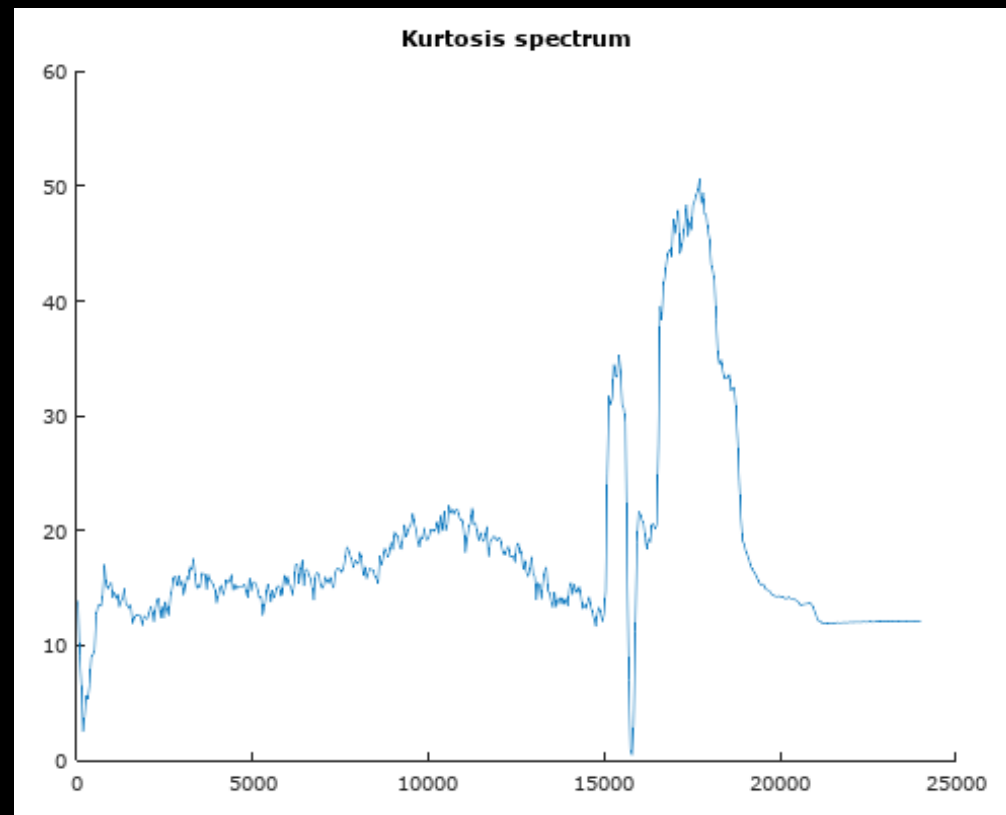
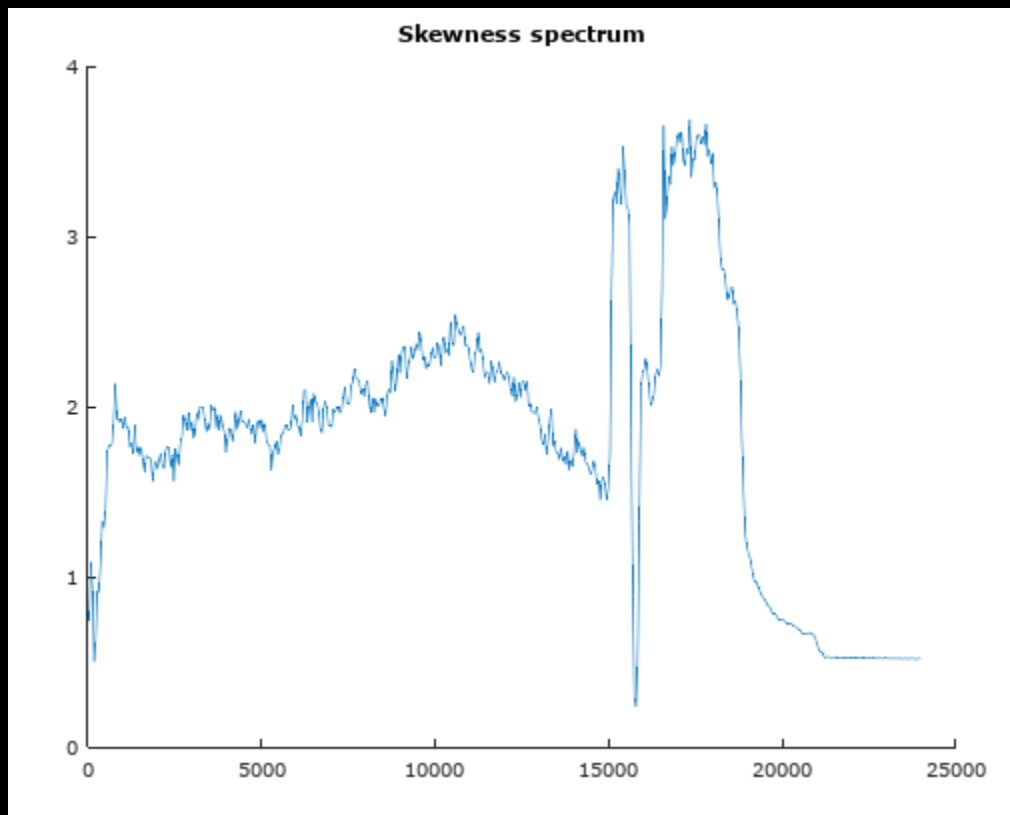
Asymétrie



Aplatissement



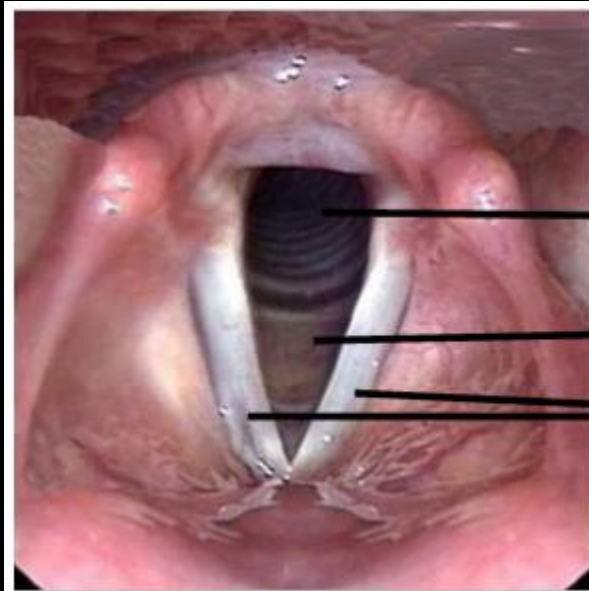
# Analyse du bruit



# Analyse de la voix

- L'analyse de la voix se base sur le fait que les propriétés acoustiques des sons sont influencées :
  - par la forme et la taille des cordes vocales,
  - et des voies respiratoires supérieures,
  - ainsi que par la position de la langue et l'ouverture de la bouche.
- Elles portent principalement sur l'analyse des voyelles.

# Analyse de la voix

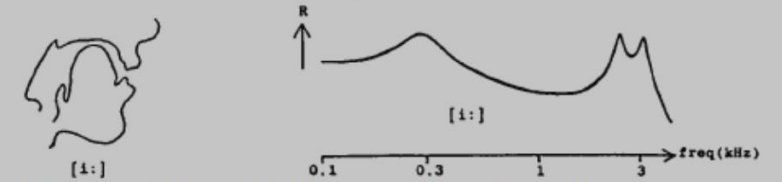


Trachée

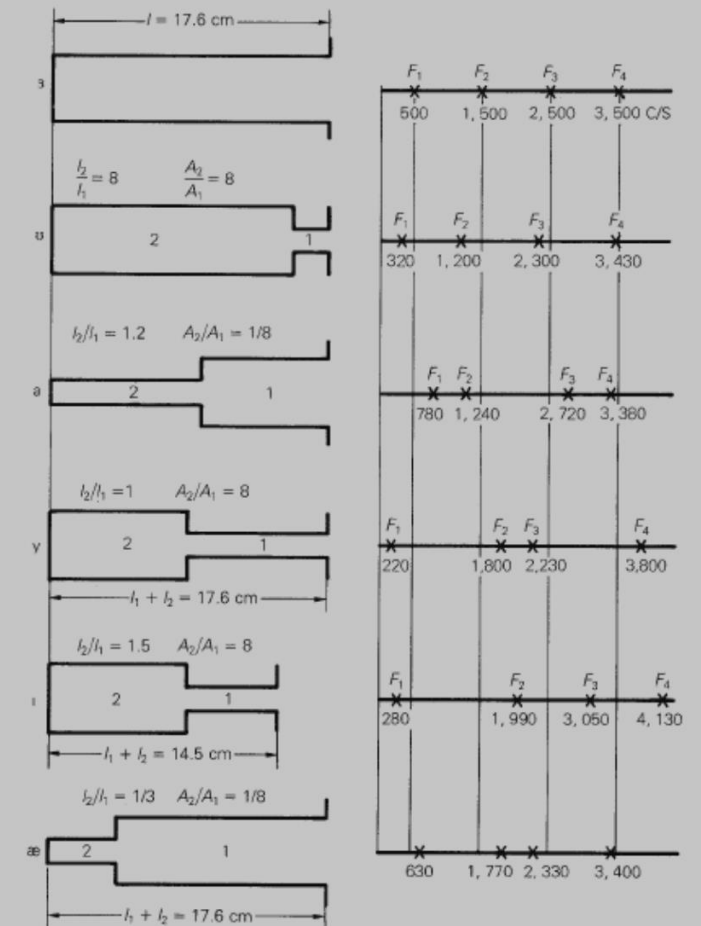
Glotte (espace vide)

Cordes vocales

Figure 2-3.3 Two Tube Resonators Approximating Vowel Articulations



What is the connection between the X-ray cross-section and the resulting frequency response?



Two-tube resonators have resonant frequencies that are very similar to vowel formant frequencies (Clark & Yallop, after Fant 1960).

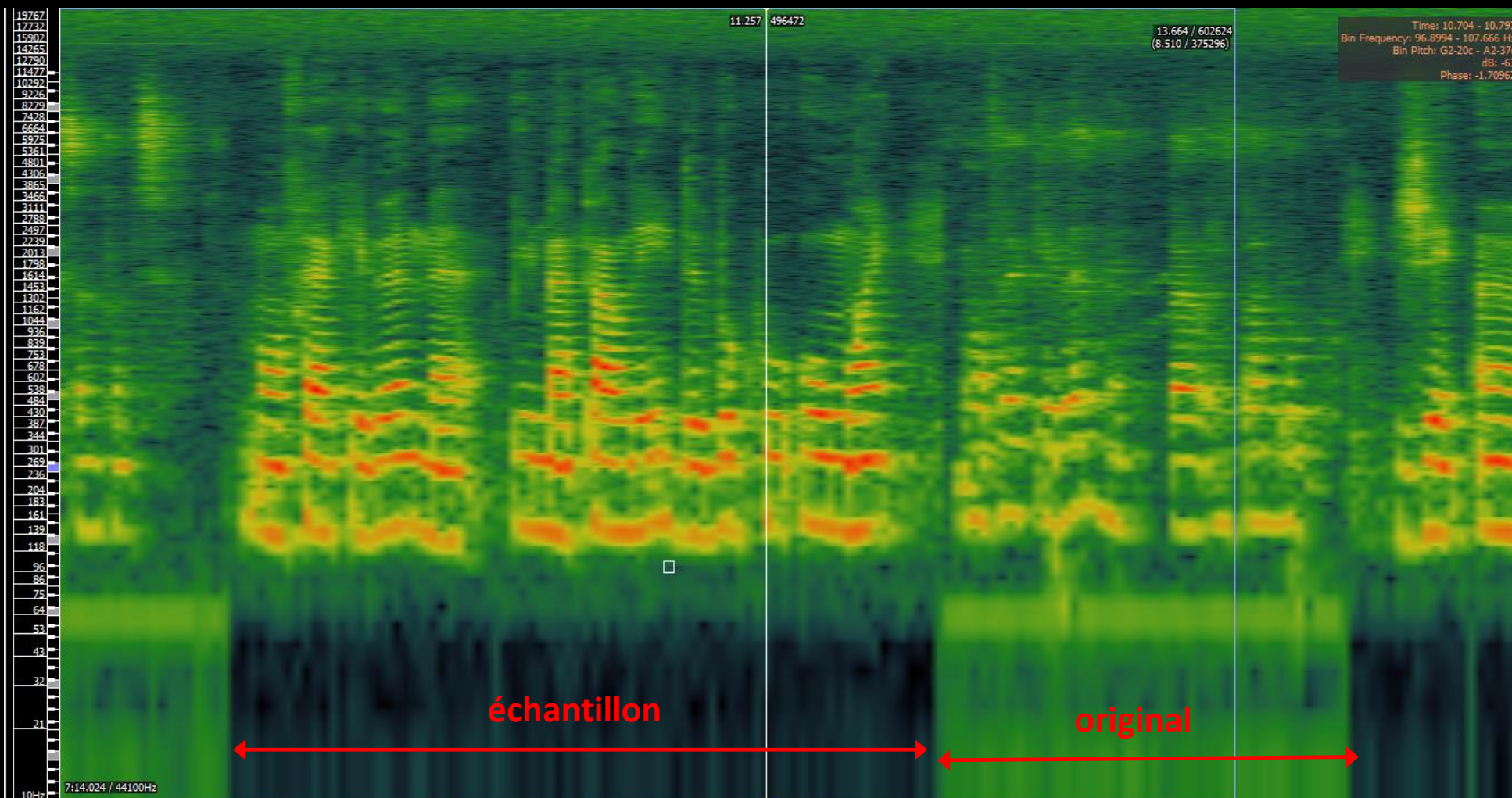
# Analyse de la voix

- *Comparaison spectro-aurale des enregistrements*
  - Méthode classique
  - À l'oreille
  - Spectrogramme
- **AMERICAN BOARD of RECORDED EVIDENCE - VOICE COMPARISON STANDARDS (1999)**
  - *There must be at least (10) comparable words between two (2) voice samples to reach a minimal decision criteria*

# AMERICAN BOARD of RECORDED EVIDENCE -- VOICE COMPARISON STANDARDS (1999)

- **Identification.** At least 90% of all the comparable words must be very similar aurally and spectrally, producing not less than **twenty (20) matching words**. Each word must have **three (3) or more usable formants**. This confidence level is not allowed when there is obvious voice or electronic disguise in either sample, or the samples are more **than six (6) years apart**.
- **Probable Identification.** At least 80% of the comparable words must be very similar orally and spectrally, producing not less than **fifteen (15) matching words**. Each word must have two (2) or more usable formants.
- **Possible Identification.** At least 80% of the comparable words must be very similar orally and spectrally, producing not less than **(10) matching words**. Each word must have two (2) or more usable formants.
- **Inconclusive.** Falls below either the Possible Identification or Possible Elimination confidence levels and/or **the examiner does not believe a meaningful decision is obtainable due to various limiting factors**. Comparisons that reveal aural similarities and spectral differences, or vice versa, must produce an Inconclusive decision.
- **Possible Elimination.** At least 80% of the comparable words must be very dissimilar orally and spectrally, producing not less than (10) that do not match. Each word must have two (2) or more usable formants.
- **Probable Elimination.** At least 80% of the comparable words must be very dissimilar orally and spectrally, producing not less than fifteen (15) words that do not match. Each word must have two (2) or more usable formants.
- **Elimination.** At least 90% of all the comparable words must be very dissimilar orally and spectrally, producing not less than twenty (20) words that do not match. Each word must have three (3) or more usable formants. This confidence level is not allowed when there is obvious voice or electronic disguise in either sample, or the samples are more than six (6) years apart.

# Analyse spectro-aurale

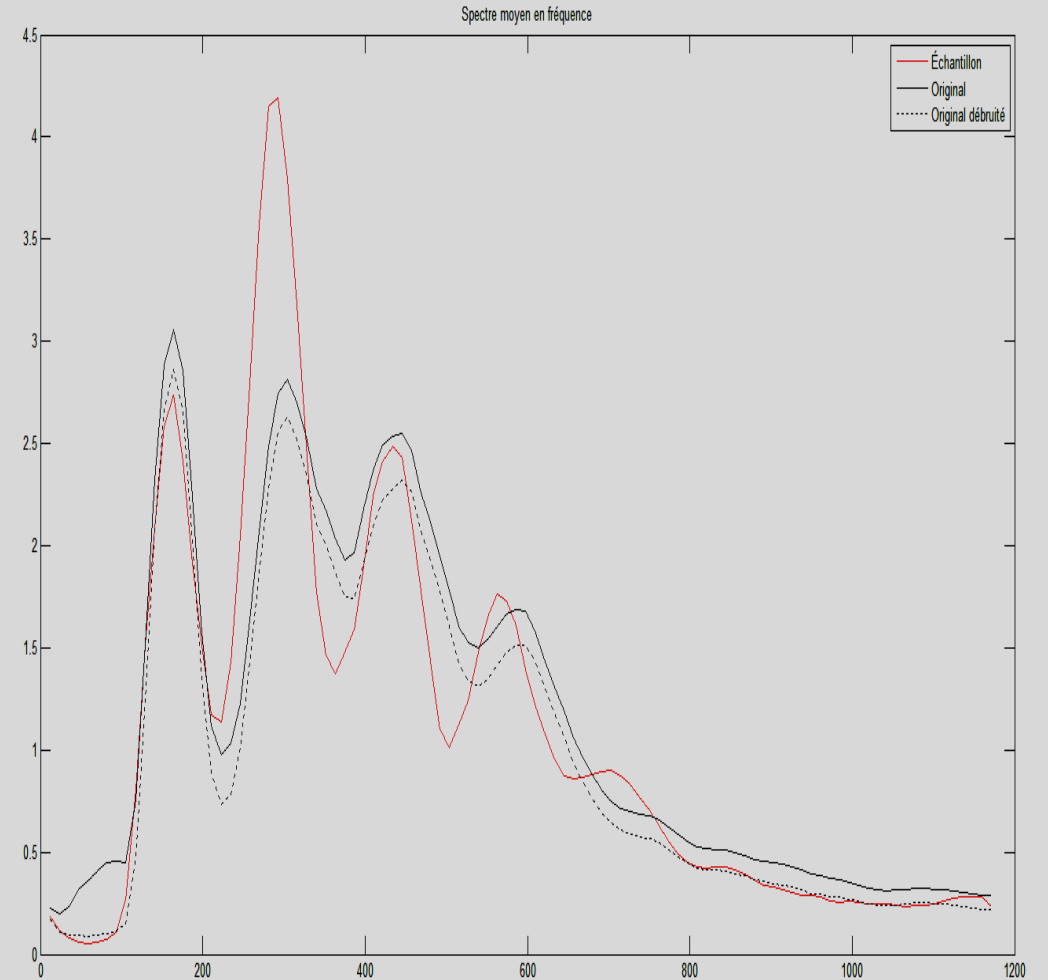


# Méthode moins subjective

- Des analyses plus quantitatives sont possibles:
  - Analyse spectrale
    - si  $> 30$  s
  - Analyse des formants
    - intéressant parce qu'interprétable physiquement
  - i-vecteur
    - Outil courant en identification de la voix, mais difficile à interpréter.
- Permettent de **quantifier** la similarité, mais demandent des banques de données pour être vraiment utiles.
  - **Analyse bayésienne**

# Spectre moyen

| Analyse du spectre moyen |                |           |                |           |                   |           |
|--------------------------|----------------|-----------|----------------|-----------|-------------------|-----------|
| Pic                      | Échantillon    |           | Original       |           | Original débruité |           |
|                          | Fréquence [Hz] | Amplitude | Fréquence [Hz] | Amplitude | Fréquence [Hz]    | Amplitude |
| 1                        | 164,1          | 2,73      | 164,1          | 3,05      | 164,1             | 2,86      |
| 2                        | 293,0          | 4,19      | 304,7          | 2,81      | 304,7             | 2,63      |
| 3                        | 433,6          | 2,49      | 445,3          | 2,55      | 445,3             | 2,32      |
| 4                        | 562,5          | 1,76      | 585,9          | 1,69      | 597,7             | 1,52      |

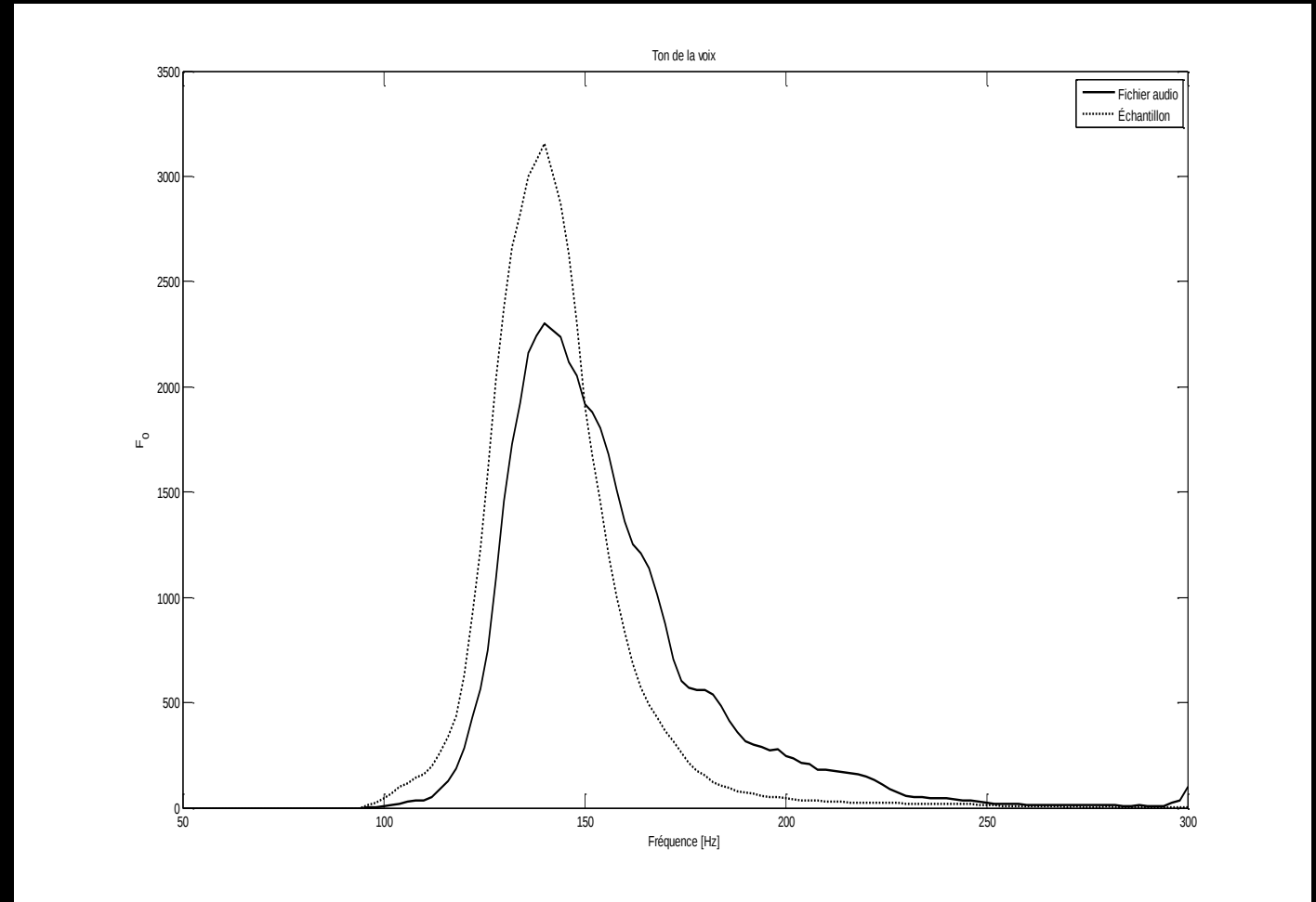


# Analyse des formants

- Les formants sont les fréquences principales de la voix humaine. Ils sont largement contrôlés par la forme des cordes vocales, des voies respiratoires et la position de la bouche.
- $F_0$  : plus basse fréquence correspond au ton de la voix.
- L'intensité et la fréquence des formants de fréquences supérieures ( $F_1$ - $F_4$ ) définissent les différentes voyelles, l'accent et le timbre de la voix.
  - $F_1$  et  $F_2$  sont suffisants pour reconnaître les voyelles
  - $F_3$  et supérieurs sont plus sensibles à aux accents et au timbre de la voix.

# Analyse des formants

- $F_0$  : ton de la voix.
- Grande similarité
- **Attention à :**
  - l'effet Lombard
  - l'effet du stress
  - l'effet de l'âge
- Facilement manipulable!



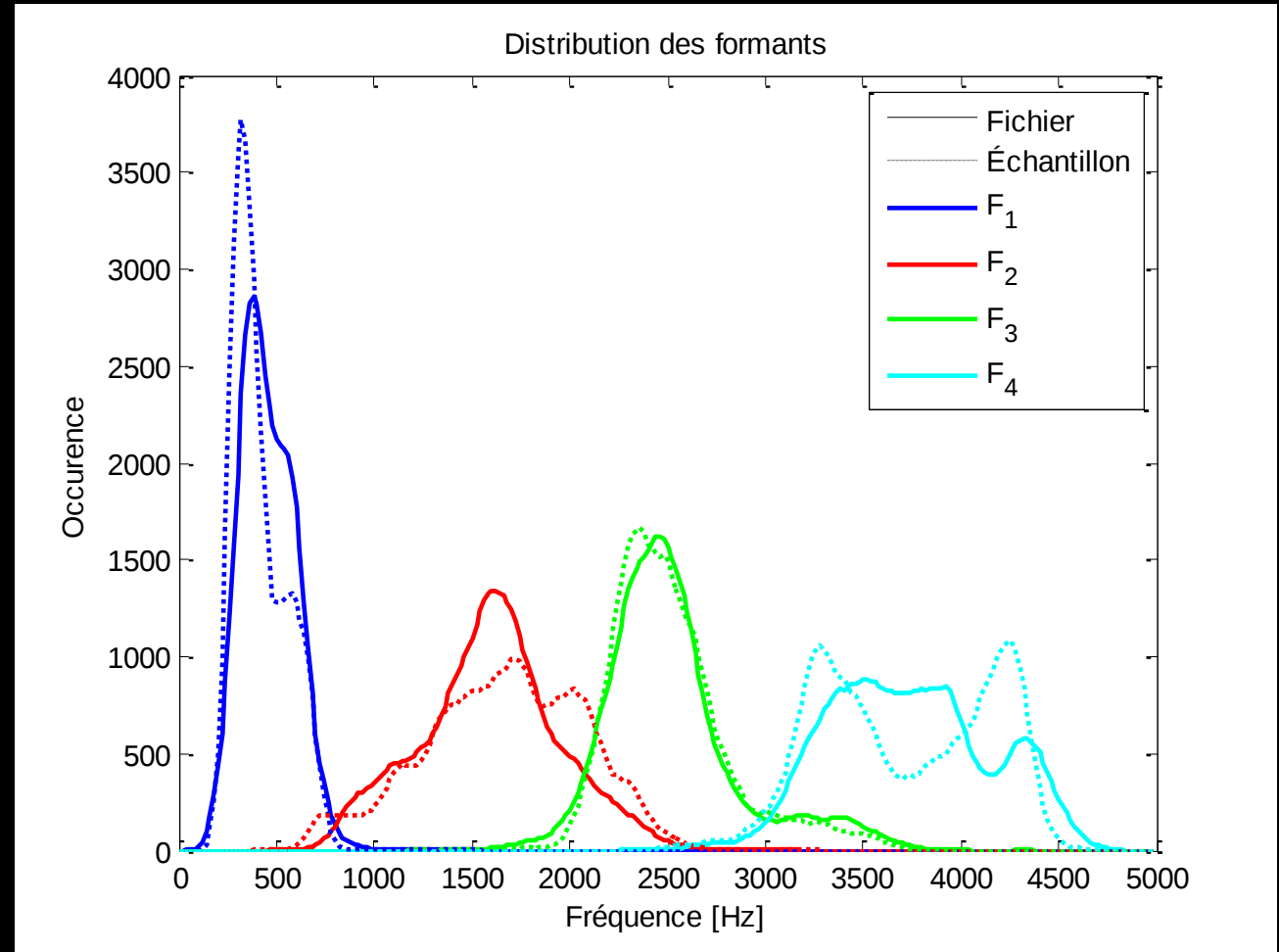
# Analyse des formants

$F_1$  et  $F_2$  sont plus facilement manipulables

- Les plus faciles à mesurer

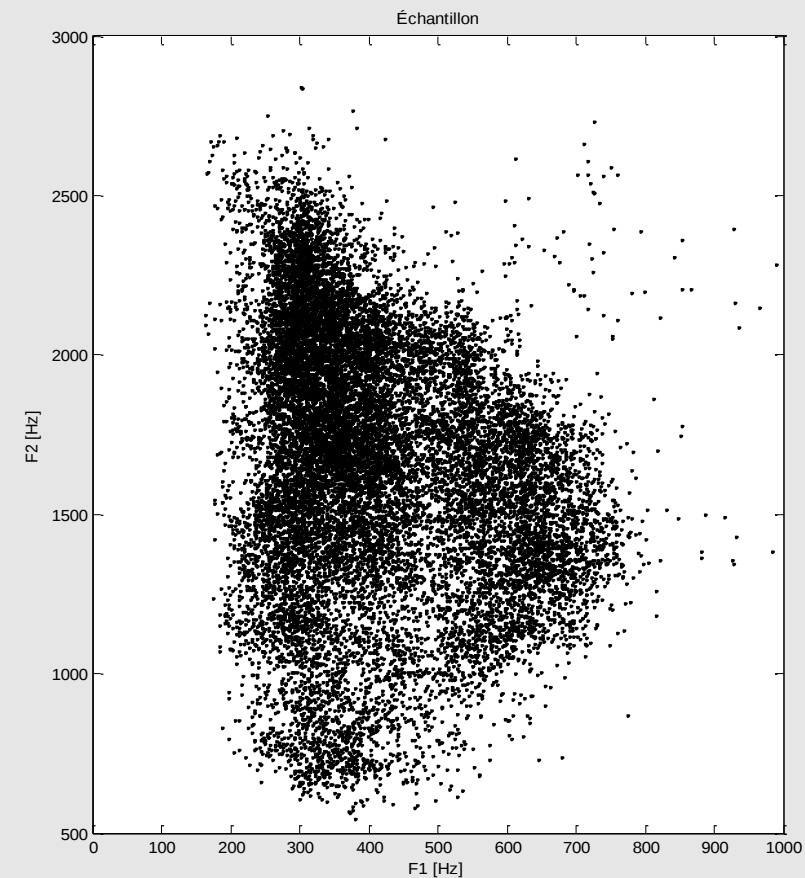
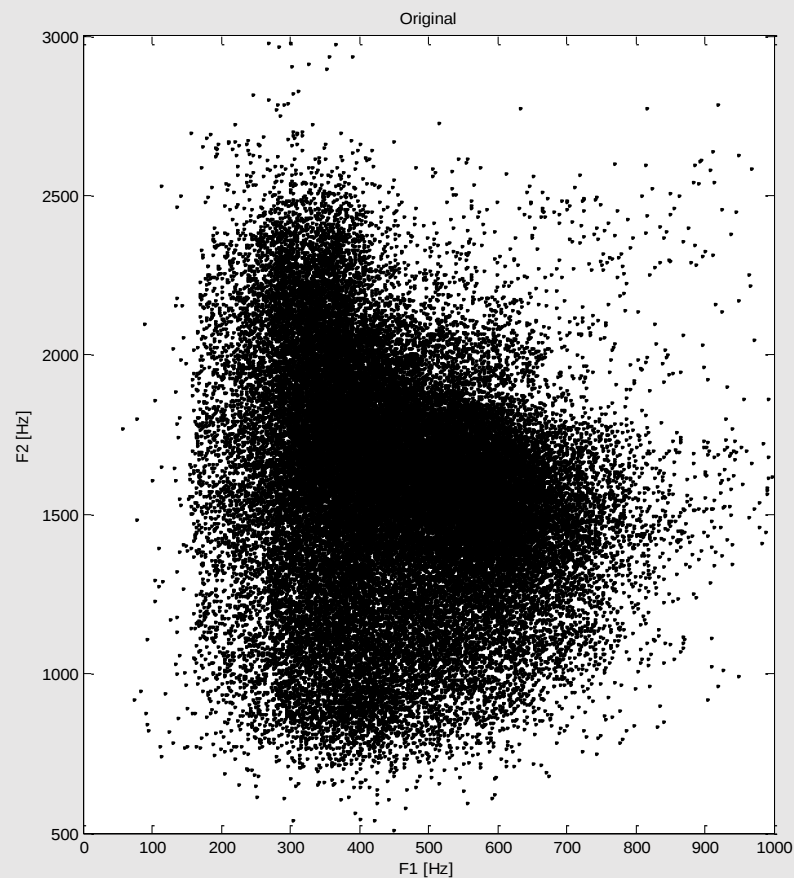
$F_3$  et  $F_4$  sont plus robustes

- Souvent bruités

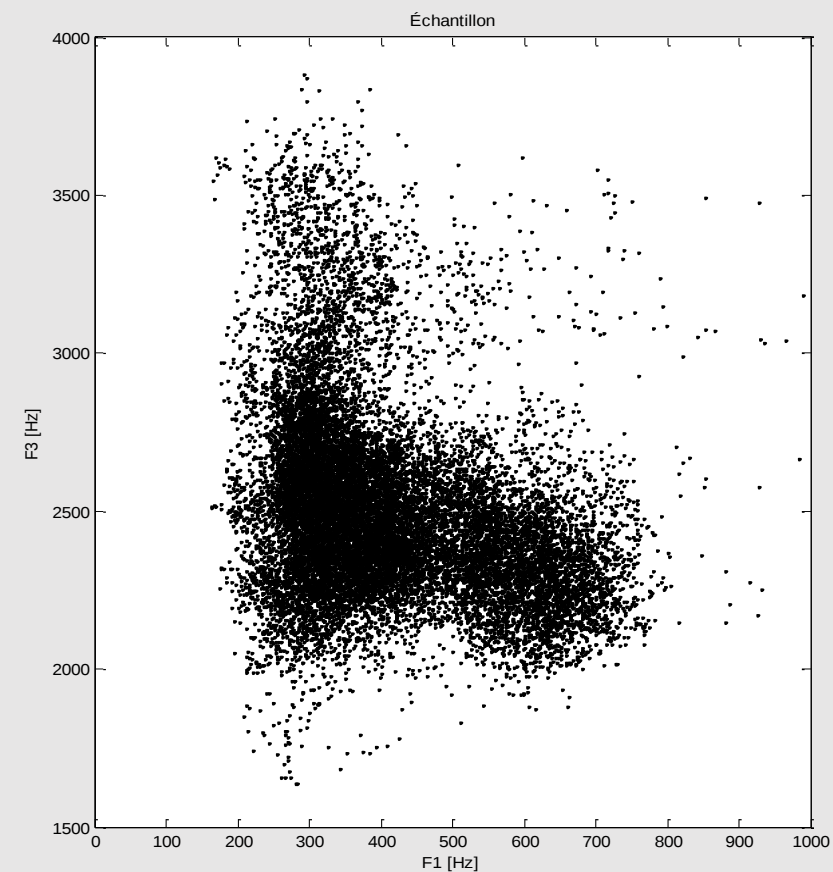
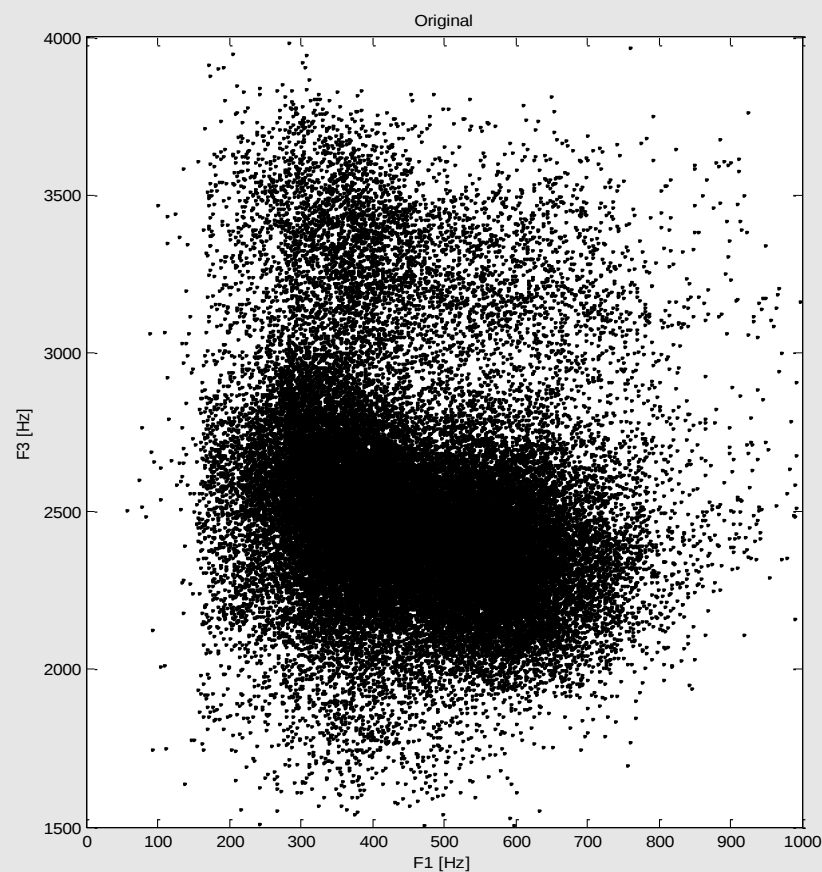


# Analyse des formants

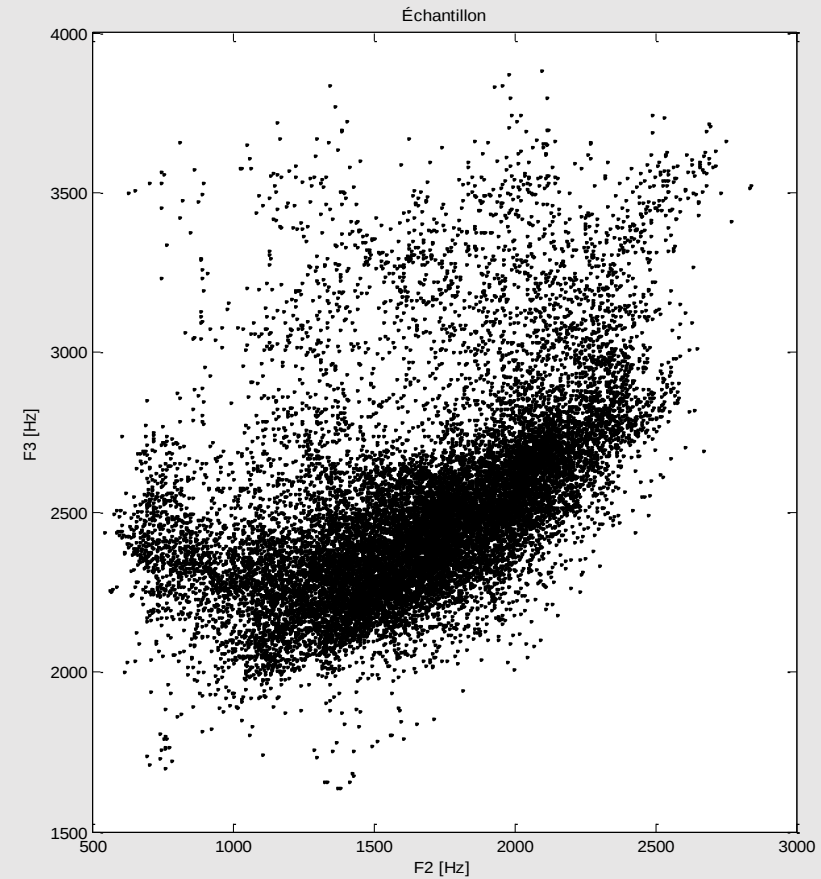
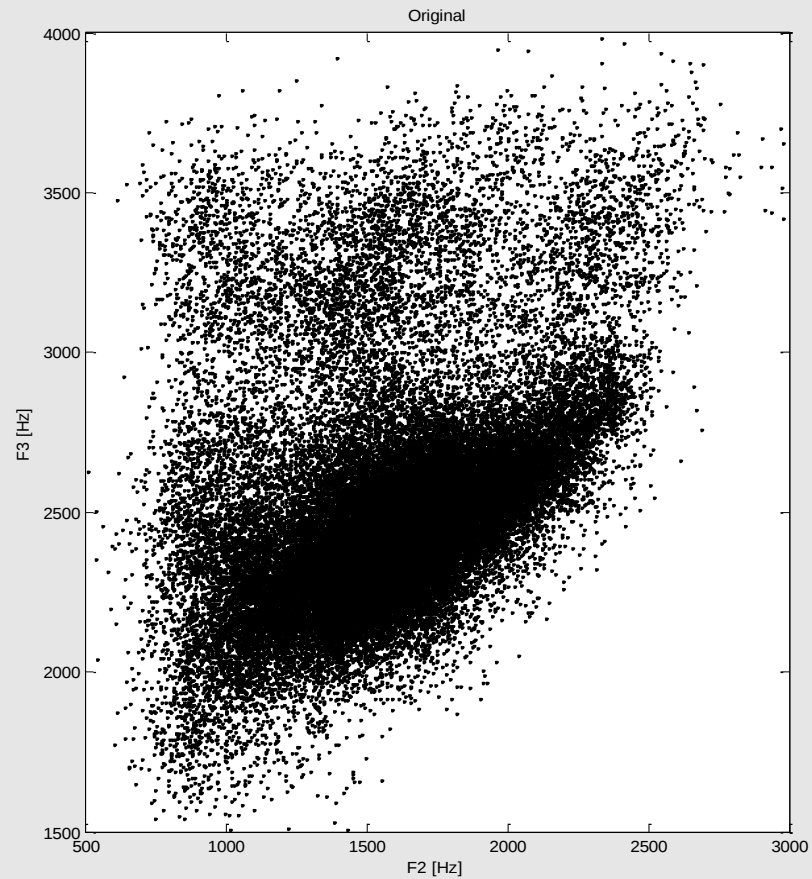
Triangle phonétique



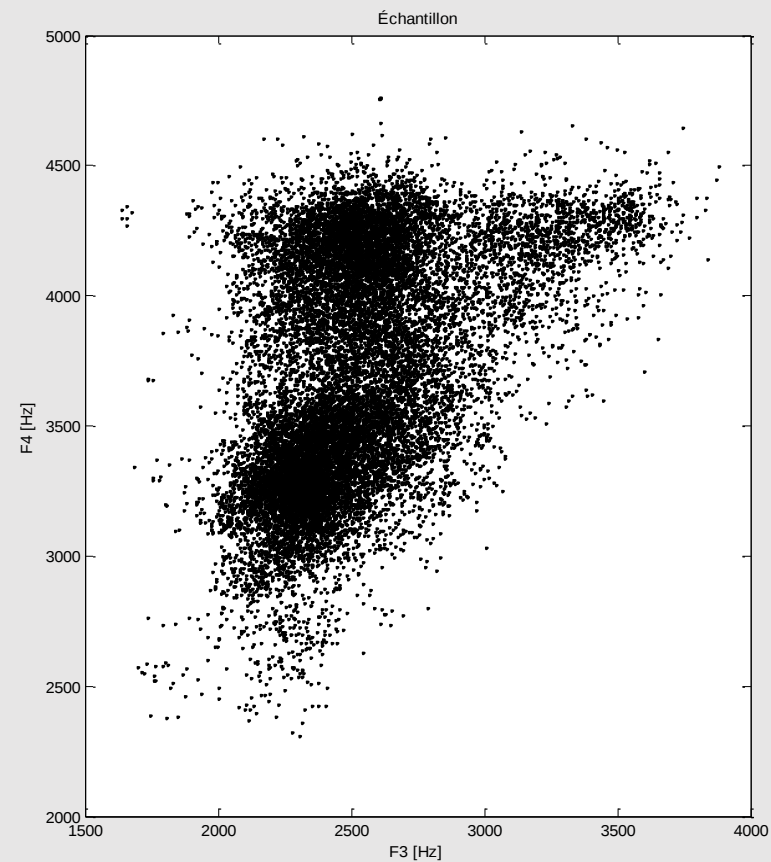
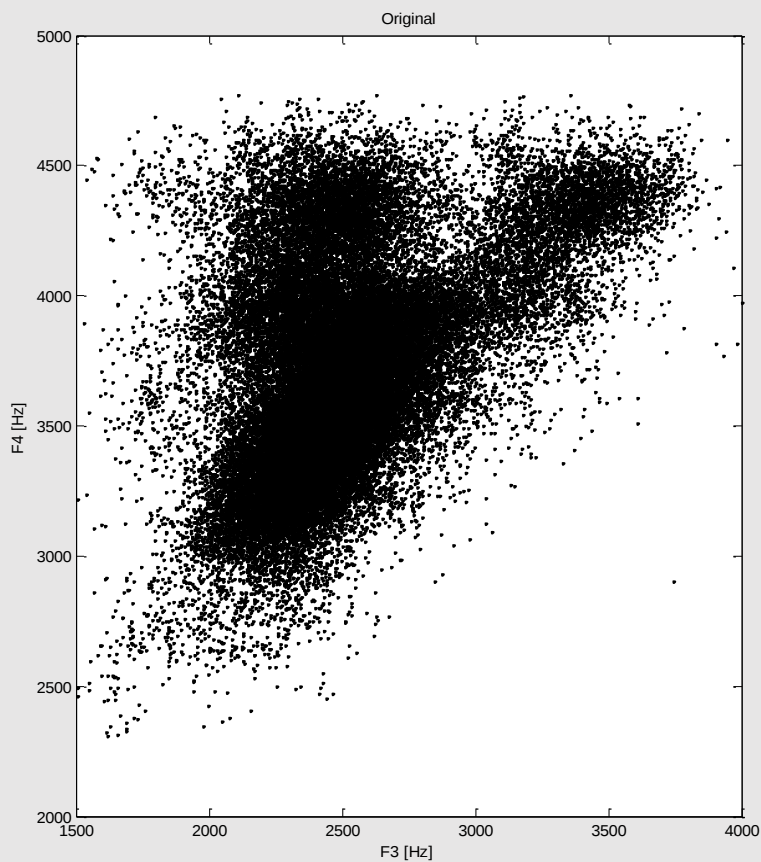
# Analyse des formants



# Analyse des formants



# Analyse des formants



# Conclusion

Au final, le résultat est ambigu

- Absence de preuve de manipulations malgré des indications
  - Manque de support de preuve matérielle
- Absence de confirmation de l'identité de la voix
  - Résultat ambigu de l'examen spectro-aural.
    - *Enregistrement de référence déficient.*
- Outils d'analyse adéquats, mais absence de population de référence.