

## DSP - Project 3

### Task 1

Έχοντας τα 3 ηχητικά σήματα<sup>1</sup> που ηχογραφήθηκαν για προηγούμενες ασκήσεις (φωνήματα /a/, /o/ και /e/), έγιναν τα εξής βήματα:

1. Για κάθε σήμα εφαρμόστηκε ο **FFT** (Fast Fourier Transform) σε ένα μέρος του (αποκοπή ενός μέρους των δειγμάτων από τη μέση του σήματος)
2. Υπολογίστηκε το **magnitude (polar representation)**:

$$\text{Mag}_{\text{signal}} = \sqrt{\text{fft\_signal}_{\text{real}}^2 + \text{fft\_signal}_{\text{imaginary}}^2}$$

3. Κρατώντας μόνο το πρώτο μισό των τιμών του magnitude, μέσα από μια επαναληπτική διαδικασία με χρήση κώδικα, σχεδιάζουμε τα **4 σημεία / bins με τη μεγαλύτερη ενέργεια**, όπως φαίνεται και στα **Figures 1, 2 και 3**.

Λόγω της φύσης των παραπάνω ηχητικών σημάτων ως **σύνθετοι** ήχοι (ανθρώπινη φωνή σε μη - επαγγελματική ηχογράφηση) και όχι καθαροί τόνοι, το γράφημα που αναπαριστά το **magnitude** των 4 πιο “ισχυρών” ενεργειακά bins (**peaks**), δεν αποτελεί τον καταλληλότερο τρόπο εύρεσης της θεμελιώδους συχνότητας αυτών των σημάτων. Ο ήχος που αποτελεί φυσική ομιλία, ξεκινά από τους παλμούς των φωνητικών μας **χορδών**, οι οποίοι καθορίζουν την  $f_0$  (fundamental frequency) του και εξέρχεται από το φωνητικό κανάλι, περνώντας από φιλτράρισμα ανάλογα με τις θέσεις και το σχήμα των **αρ-θρωτών**, αποκτώντας και **αρμονικές** συχνότητες (formants λόγω αντίκλισης στην κοιλότητα). Αποτελεί **ψυχοακουστικό** χαρακτηριστικό, καθώς αντίληψή της διέπεται από **υποκειμενικότητα**, η οποία εξαρτάται πολλές φορές από διάφορα άλλα χαρακτηριστικά όπως η **ένταση** ή η **φυσιολογία** του ανθρώπου που την αντιλαμβάνεται. Στα γραφήματα που ακολουθούν στο πεδίο των **συχνοτήτων**, θα δούμε πως η θεμελιώδης συχνότητα (**pitch**) ενός ηχητικού σήματος, δεν αποτελεί πάντα την κορυφή με τη μεγαλύτερη **ενέργεια**, καθώς υπάρχουν και οι **αρμονικές** αυτής, που βρίσκονται σε συχνότητες με τιμή, ακέραιο πολλαπλάσιο της  $f_0$  (κυρίως components  $f_1$  και  $f_2$ ). Επιπλέον, σε περιπτώσεις που το input sample του αλγορίθμου FFT δεν είναι ακέραιο πολλαπλάσιο της περιόδου του, παρατηρείται το φαινόμενο “**spectral leakage**”, παρουσιάζοντας παραποιημένες πληροφορίες για την  $f_0$ . Αντ’ αυτού, υπάρχουν άλλες μέθοδοι όπως για παράδειγμα ο αλγόριθμος “**autocorrelation**” (αυτοσυσχέτισης), που είναι αρκετά αποδοτικός σε περιπτώσεις σημάτων με μεγάλη περιοδικότητα, όπως τα ηχητικά σήματα της άσκησης (έχει αναπτυχθεί κώδικας για την εύρεση του pitch μέσω της συνάρτησης **librosa.autocorrelation()**).

Το αποτέλεσμα της εφαρμογής της συνάρτησης **autocorrelation()** της βιβλιοθήκης **librosa** ως προς τη θεμελιώδη συχνότητα των σημάτων, ήταν τα παρακάτω:

- $f_0$  του σήματος με το φώνημα / a / : **233.33 Hz**
- $f_0$  του σήματος με το φώνημα / o / : **219.40 Hz**
- $f_0$  του σήματος με το φώνημα / e / : **297.97 Hz**

<sup>1</sup> Χρησιμοποιήθηκαν μόνο τα **φωνήεντα**, καθώς τα σύμφωνα που είχαν επιλεγεί ανήκουν στην κατηγορία των μη ηχηρών φθόγγων, των οποίων τα σήματα δεν έχουν περιοδικότητα συνεπώς ούτε και θεμελιώδη συχνότητα, εφόσον κατά την εκφώνησή τους δεν πάλλονται οι φωνητικές χορδές.

Αξιοποιώντας το λογισμικό Praat, έγιναν επιπλέον μετρήσεις για τα formants και την  $f_0$  των σημάτων (mean values):

- φώνημα / **a** /
  - $f_0$  : ~ **233 Hz**
  - $f_1$  : ~ **706 Hz** (αν και αποτελεί 3η αρμονική)
- φώνημα / **o** /
  - $f_0$  : ~ **219 Hz**
  - $f_1$  : ~ **490 Hz**
  - $f_2$  : ~ **920 Hz**
- φώνημα / **e** /
  - $f_0$  : ~ **298 Hz**
  - $f_1$  : ~ **560 Hz**

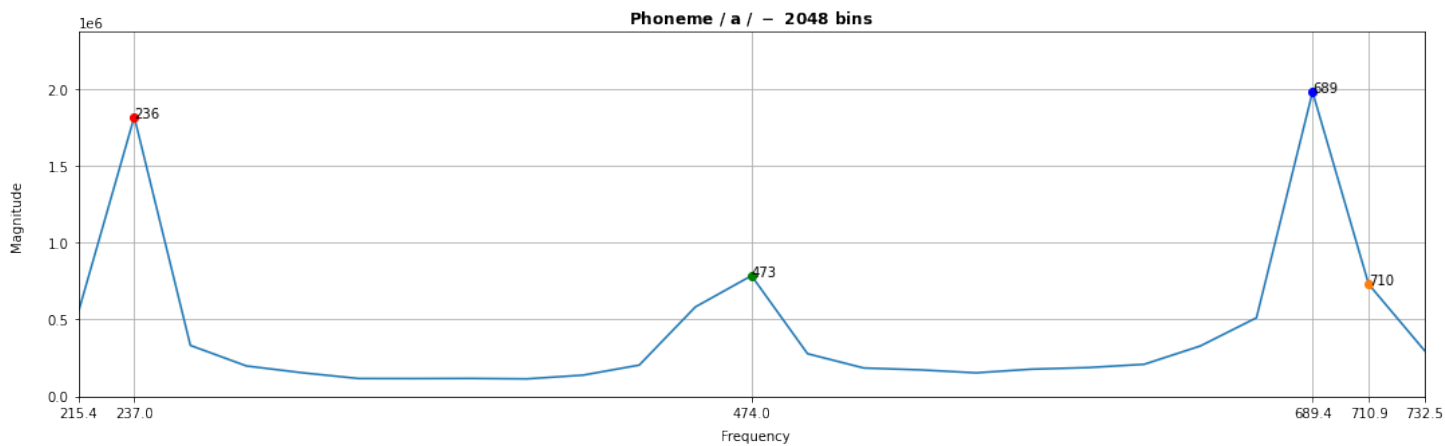
Τα σήματα των φωνημάτων **a**, **o** και **e**, είχε υπολογιστεί “με το μάτι” πως είχαν αντίστοιχα συχνότητες ~**230 Hz**, ~**217 Hz** και ~ **294 Hz** , συνεπώς υπάρχει εμφανής **απόκλιση** (αν και μικρή) μεταξύ των τιμών από τους παραπάνω υπολογισμούς και εκείνων της προηγούμενης εργασίας (σημαντικός παράγοντας αποτελεί και το σημείο από το οποίο προερχόταν το δείγμα της προηγούμενης εργασίας, που ίσως, όπως αποδείχθηκε, να μην ήταν το καταλληλότερο για τις μετρήσεις “με το μάτι”).

Τα γραφήματα **1**, **2** και **3**, μας οδηγούν στα εξής συμπεράσματα:

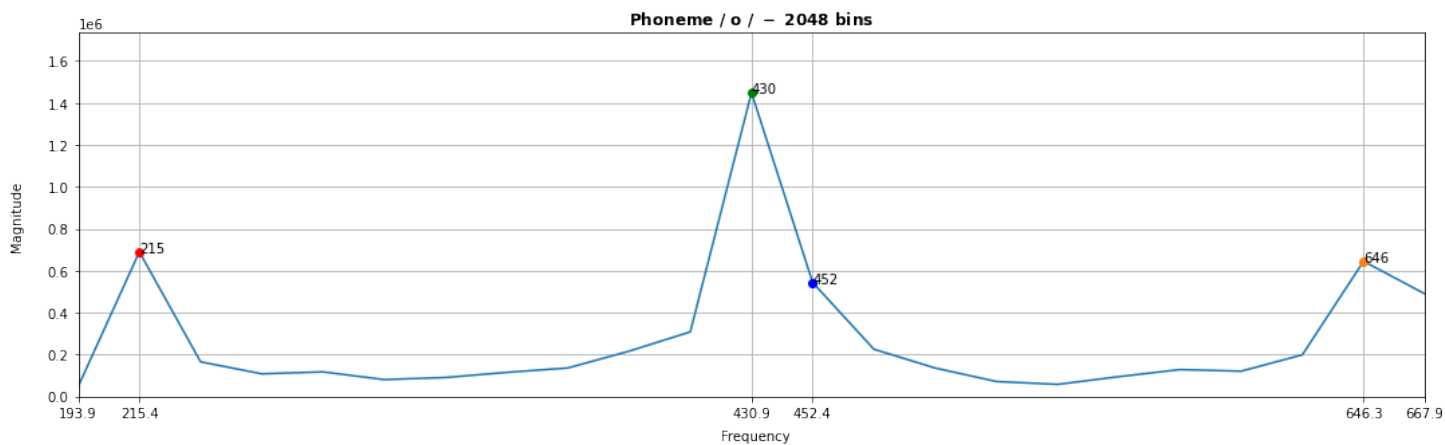
- Σε όλες τις περιπτώσεις εντοπίζεται η θεμελιώδης συχνότητα ή έστω μια προσεγγιστική τιμή αυτής
- Παρουσιάζονται ακέραια πολλαπλάσια της  $f_0$  (ή προσέγγιση της τιμής τους) και κυρίως οι  $f_1$  και  $f_2$  (αρμονικές συχνότητες)
- Η υψηλότερη κορυφή (peak) άρα το σημείο με τη μεγαλύτερη “ενέργεια”, δεν είναι πάντοτε εκείνο που αντιστοιχεί στο bin της θεμελιώδους συχνότητας, καθώς όπως βλέπουμε, μπορεί η κορυφή να αντιστοιχεί σε αρμονική συχνότητα αυτής

Τελικά, όπως έχουν δείξει μελέτες, η θεμελιώδης συχνότητα απομονωμένα, δεν παίζει τόσο σημαντικό ρόλο στο task της **αναγνώρισης ομιλίας** (φωνημάτων). Το **φασματογράφημα** από την άλλη, τόσο σαν πηγή ποσοτικών δεδομένων μετά από μετρήσεις, όσο και από την ίδια την εικόνα του, μπορεί να προσφέρει αρκετές πληροφορίες για τα φωνήματα που έχουν εκφωνηθεί και αναπαρασταθεί σε αυτό.

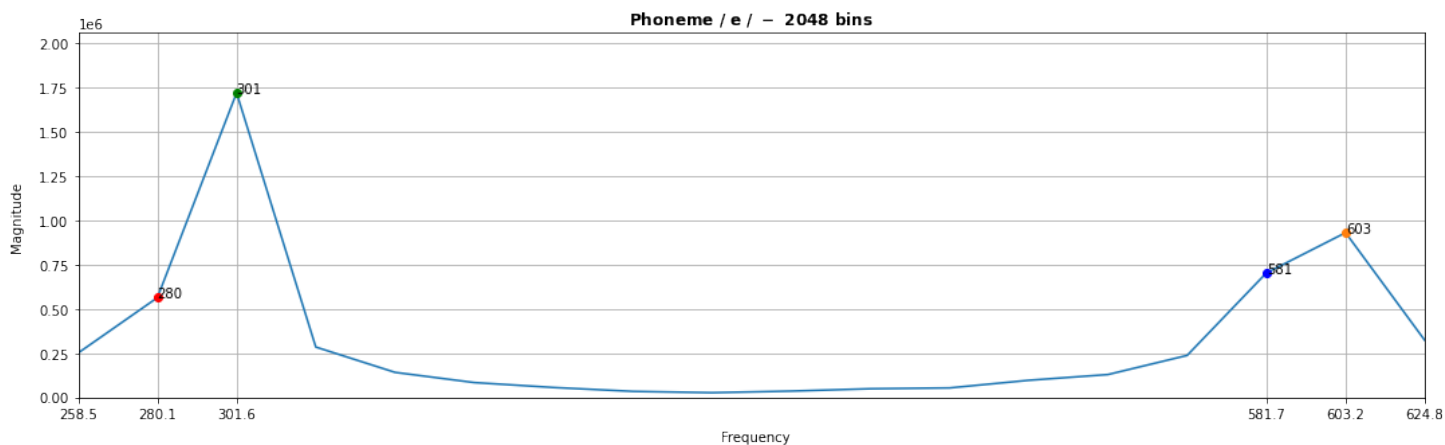
Ακολουθούν τα **3 γραφήματα**.



**Figure 1**



**Figure 2**



**Figure 3**

## Task 2

Αρχικά, δημιουργήθηκε το τεχνητό σήμα

$$s = s_1 + s_2$$

όπου

$$s_1 = \sin(2\pi f_1 t), \quad f_1 = 50\text{Hz}$$

$$s_2 = \sin(2\pi f_2 t), \quad f_2 = 70\text{Hz}$$

με συχνότητα δειγματοληψίας  $f_s = 44100\text{Hz}$ .

Στη συνέχεια εφαρμόστηκε σε αυτό ο αλγόριθμος **FFT**.

Με παρόμοια λογική και διαδικασία όπως και στο **Task 1**, εντοπίστηκαν τα κοντινότερα bins στις συχνότητες του σήματος  $s$  και σχεδιάστηκαν σε γραφήματα, με διαφορετικό αριθμό **bins** κάθε φορά, ώστε να παρατηρηθεί το **ελάχιστο πλήθος** αυτών, ώστε να βρίσκονται οι 2 συχνότητες  $f_1, f_2$  σε διαφορετικά bins.

Το εύρος συχνοτήτων που αντιπροσωπεύει κάθε bin, βρίσκεται ως εξής:

$$\text{resolution}_{\text{bin}} = \frac{f_s}{\text{window\_size}} \quad (1)$$

Αρα για παράδειγμα με τα δικά μας δεδομένα και με βάση τη σχέση (1) :

$$\text{window\_size} = 512 \implies \text{resolution}_{512} = \frac{f_s}{\text{window\_size}} = \frac{44100}{512} = 86,132$$

$$\text{window\_size} = 1024 \implies \text{resolution}_{1024} = \frac{f_s}{\text{window\_size}} = \frac{44100}{1024} = 43.066$$

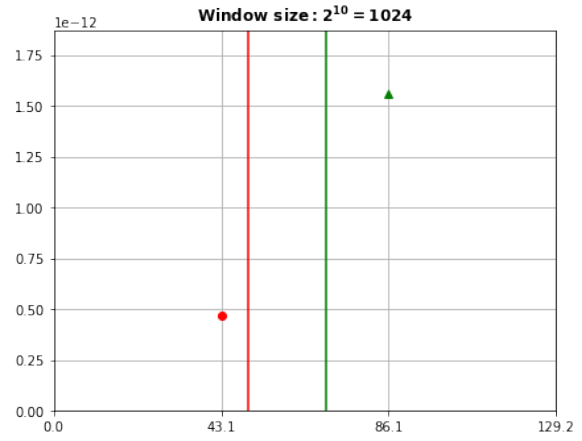
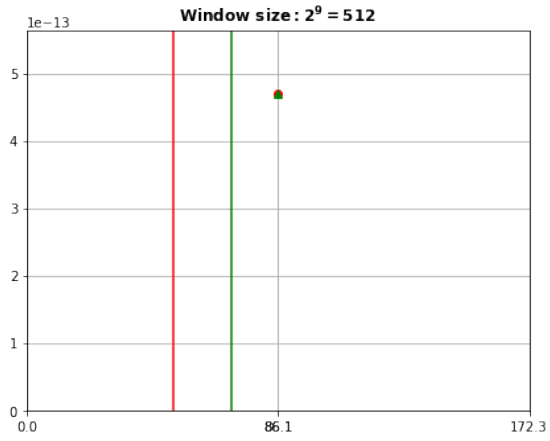
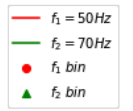
Αρα με το συγκεκριμένο  $f_s = 44100\text{Hz}$ , το ελάχιστο πλήθος bins ώστε να διαχωρίζονται οι συχνότητες 50 και 70 Hz είναι όπως φαίνεται από τις παραπάνω σχέσεις, τα **1024** bins, καθώς λόγω του resolution όταν αυτά είναι 512 σε πλήθος, το 1ο bin θα περιλαμβάνει **και τις 2** αυτές συχνότητες, ενώ στα 1024, η συχνότητα 50Hz θα βρεθεί στο 1ο bin και η συχνότητα 70Hz στο 2ο.

Αντίστοιχα, εφαρμόζοντας τις ίδιες σχέσεις και για την

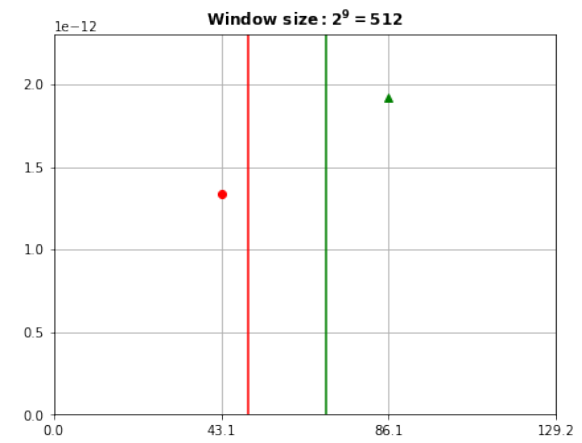
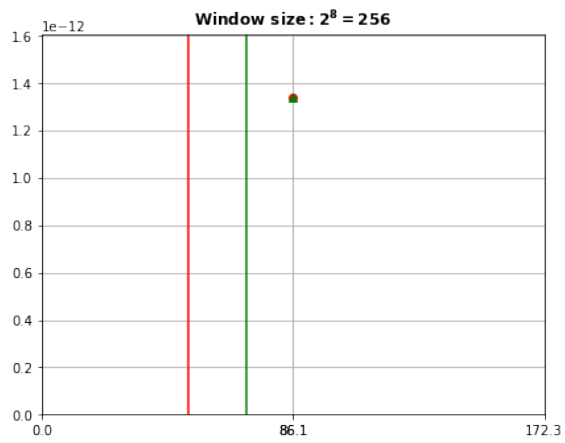
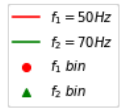
$$f'_s = \frac{f_s}{2} = \frac{44100}{2} \text{ Hz} = 22050 \text{ Hz},$$

προκύπτει πως το ελάχιστο πλήθος των bins θα είναι **512**, δηλαδή ακριβώς τα μισά από εκείνα της προηγούμενης περίπτωσης.

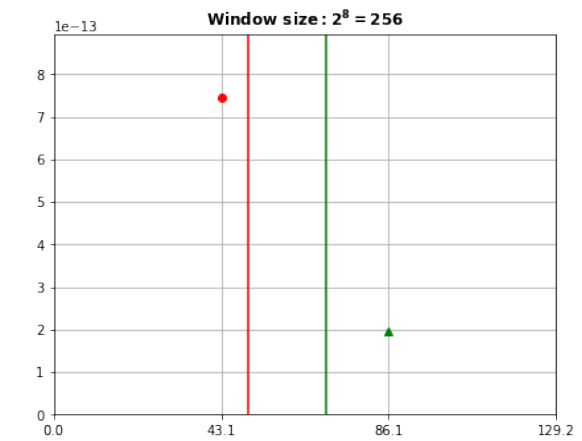
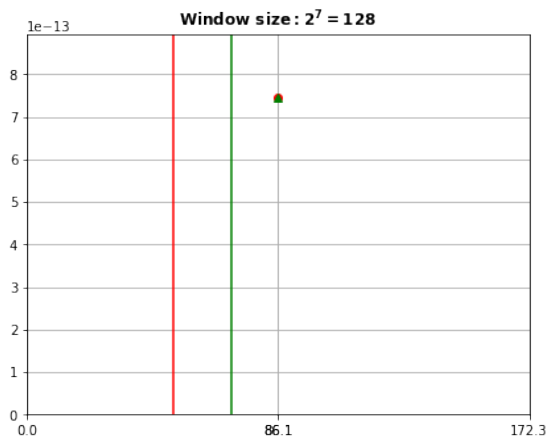
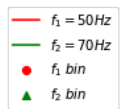
Sample Rate : 44100Hz



Sample Rate : 22050Hz

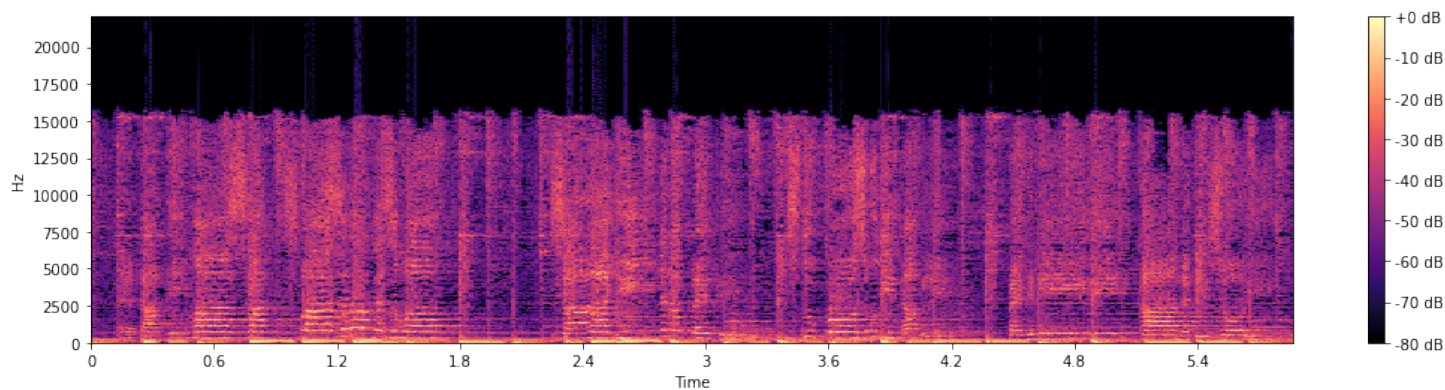


Sample Rate : 11025Hz

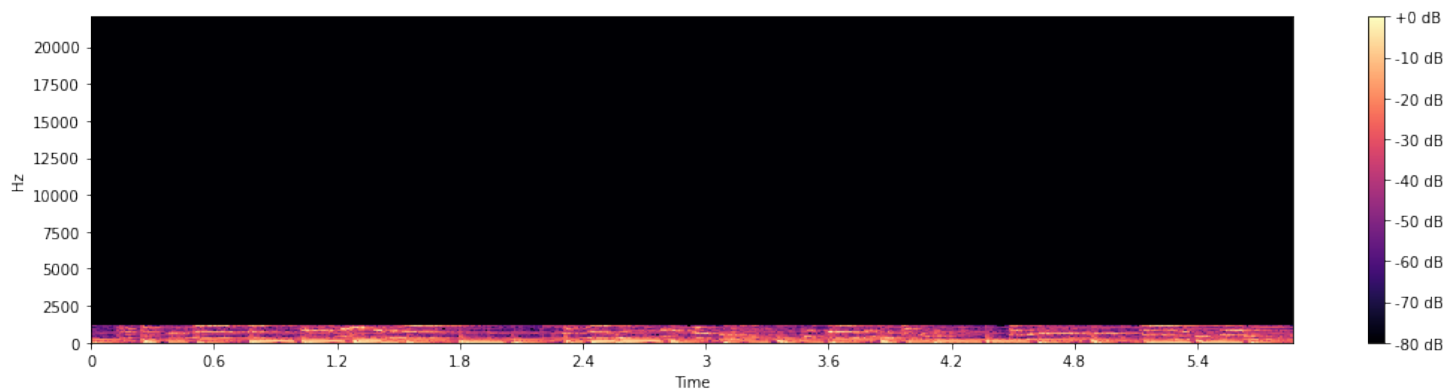


# Task 3

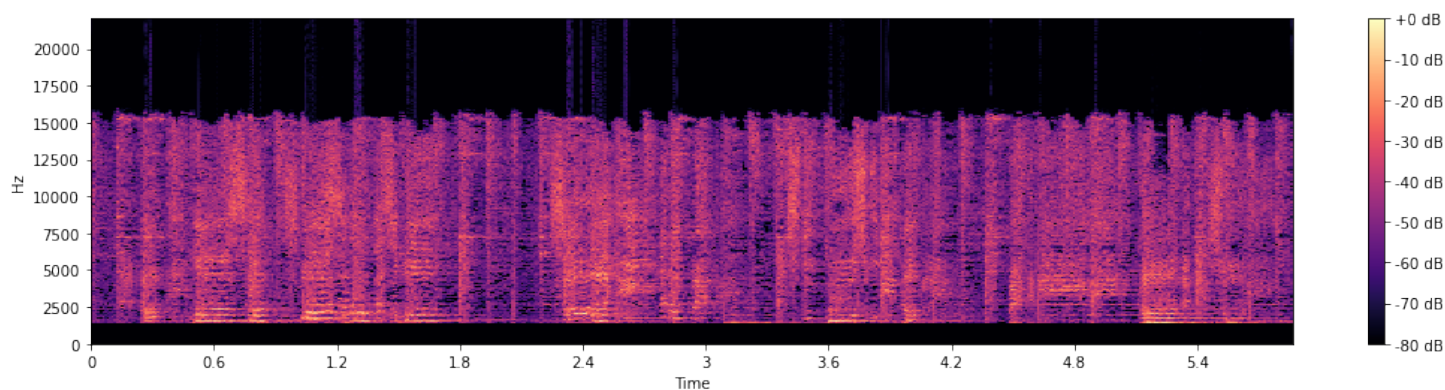
Σε αυτό το task, αξιοποιήθηκε ένα μικρό τμήμα από το τραγούδι “10 μπλε στρουμφάκια”, στο οποίο εφαρμόστηκαν 4 φίλτρα, ένα χαμηλοπερατό (low pass filter), ένα υψηλοπερατό (high pass filter) και 2 φίλτρα τυχαίας αποκοπής (πειράματα). Τα φασματογραφήματα παρουσιάζονται παρακάτω, και τα μουσικά αρχεία της εκάστοτε επεξεργασίας, παράγονται κατά την εκτέλεση του κώδικα.



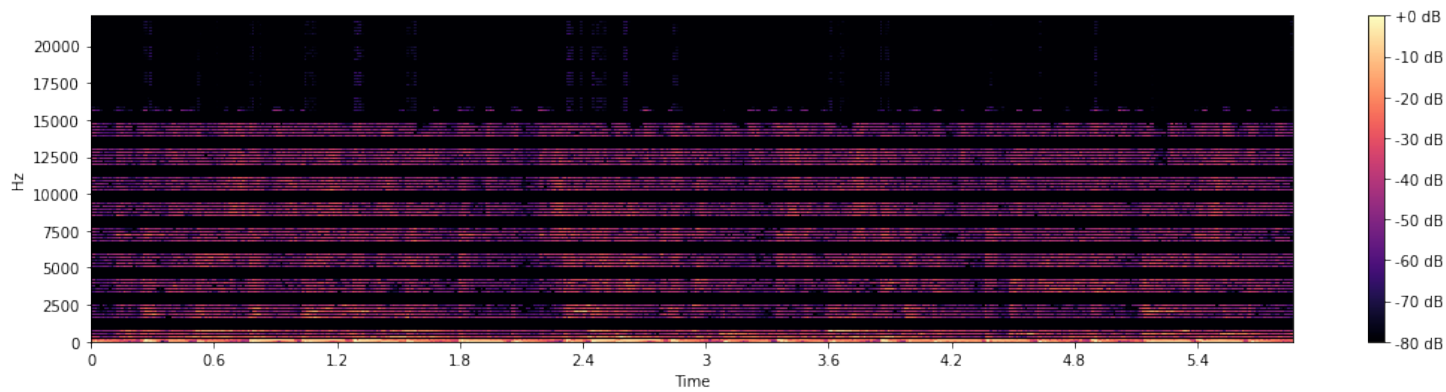
Unfiltered



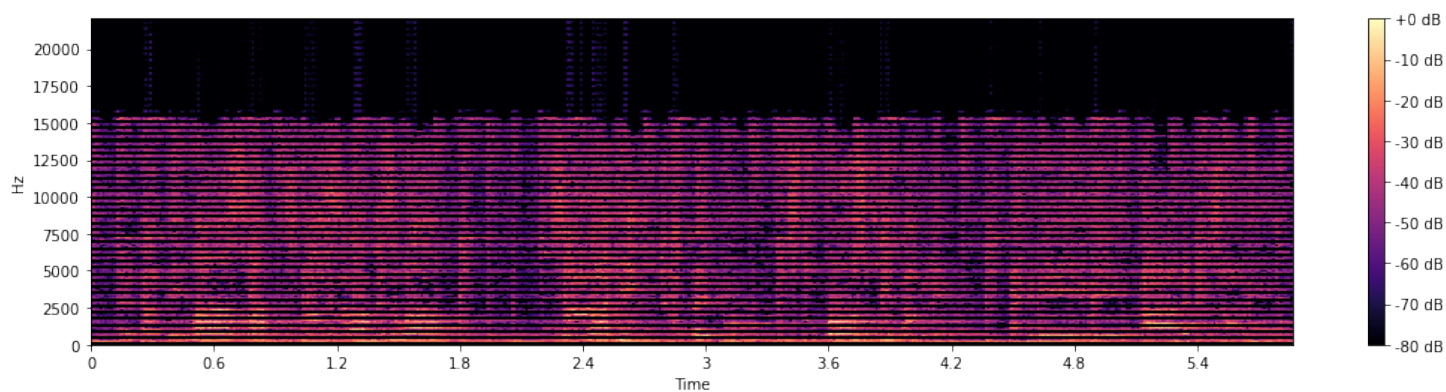
LPF



HPF



Experiment 1



Experiment 2

## Notes

Για την υλοποίηση των παραπάνω ζητουμένων, αναπτύχθηκε κώδικας που θα σταλεί μαζί με το παρόν αρχείο.