Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής
Ψηφιακή Επεξεργασία Σήματος

2η Εργαστηριακή Άσκηση
Κωδικοποίηση σημάτων μουσικής βάσει του ψυχοακουστικού μοντέλου (Perceptual Audio Coding)

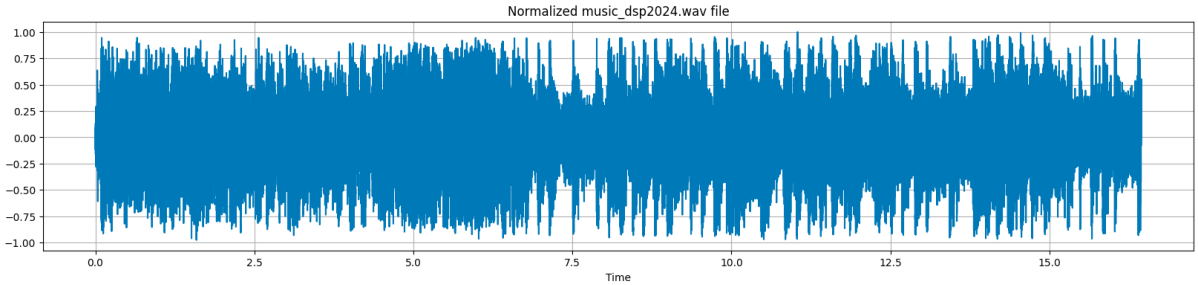
Ονοματεπώνυμο	Γεώργιος Αλέξανδρος Γεωργαντζάς	Βλάσιος Σκουλής
ΑΜ	03120017	03120065

Σάββατο 18 Μαΐου 2024

Μέρος 1: Ψυχοακουστικό Μοντέλο 1

1.0 Προεπεξεργασία του σήματος

Διαβάζουμε το αρχείο μουσικής music.wav το οποίο είναι σε μορφή stereo, δηλαδή αποτελείται από δύο channels. Για την μετατροπή του σε mono, απλά παίρνουμε τον μέσο όρο των δύο channels σε κάθε δείγμα. Στη συνέχεια κανονικοποιούμε το σήμα στο διάστημα [-1, +1] και παίρνουμε το παρακάτω διάγραμμα:



Εικόνα 1

1.1 Φασματική Ανάλυση

Ορίζουμε την συνάρτηση **bark_scale()** η οποία παίρνει ως είσοδο συχνότητα σε Hz και την επιστρέφει σε Bark σύμφωνα με τον παρακάτω τύπο:

$$b(f) = 13\arctan(.00076f) + 3.5\arctan[(f/7500)^2]\ (\text{Bark})$$

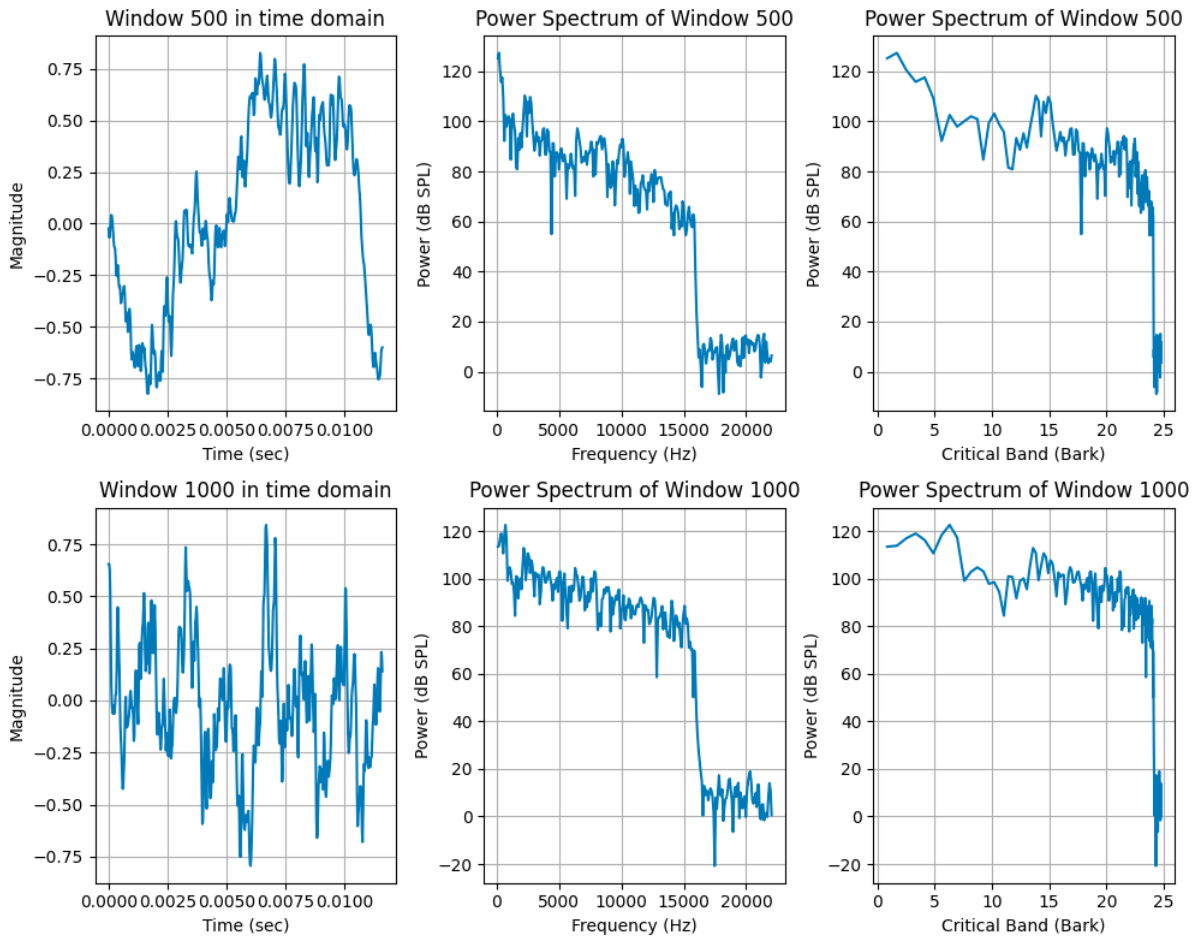
Στη συνέχεια, υπολογίζουμε το φάσμα ισχύος P(k), N = 512 σημείων, για κάθε παραθυροποιημένο τμήμα του αρχικού σήματος. Για κάθε τέτοιο τμήμα, υπολογίζεται σύμφωνα με τον παρακάτω τύπο:

$$P(k) = PN + 10\log_{10}\left|\sum_{n=0}^{N-1} w(n)x(n)e^{-j\frac{2\pi kn}{N}}\right|^2, 0 \le k \le \frac{N}{2}$$

, όπου PN = 90.302 dB και w(n) το παράθυρο hanning μήκους N = 512 σημείων:

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

Παρατηρούμε ότι ο τύπος του φάσματος ισχύος, κρύβει μέσα του τον DFT κάθε παραθυροποιημένου τμήματος. Επομένως, το P(k) έχει ανάλογη συμμετρία με αυτόν και για το λόγο αυτό κρατάμε τα πρώτα N / 2 = 256 δείγματα.
Το σήμα μας αποτελείται συνολικά από **725210** δείγματα, ενώ το διαιρούμε σε παράθυρα μήκους **512** δειγμάτων. Επειδή η διαίρεση **725210 / 512** δεν είναι τέλεια αναγκαζόμαστε να κάνουμε **zero-padding** στο σήμα μουσικής. Έτσι το χωρίζουμε τελικά σε 1417 παράθυρα. Όλοι οι υπολογισμοί που αναφέρονται στην εργασία αυτή, γίνονται για όλα τα παράθυρα του σήματος. Παρόλα αυτά, για να μπορούμε να αναπαραστήσουμε και να οπτικοποιήσουμε τους υπολογισμούς μας, διαλέγουμε τυχαία τα παράθυρα 500 και 1000, και τα χρησιμοποιούμε για διαγράμματα:



Εικόνα 2

1.2 Εντοπισμός μασκών τόνων και θορύβου (Maskers)

Προκειμένου να ανιχνεύσουμε τις διακριτές συχνότητες που αποτελούν τονικές μάσκες, χρειάζεται να ανιχνεύσουμε τις θέσεις τοπικών μεγίστων σε καθορισμένες περιοχές. Για το λόγο αυτό χρησιμοποιούμε την παρακάτω Boolean συνάρτηση $S_T(k)$:

$$S_T(k) = \begin{cases} 0, & \text{αν } k \notin [2, 250) \\ P(k) > P(k \pm 1) \bigwedge P(k) > P(k \pm \Delta_k) + 7\text{dB}, & \text{αν } k \in [2, 250) \end{cases}$$

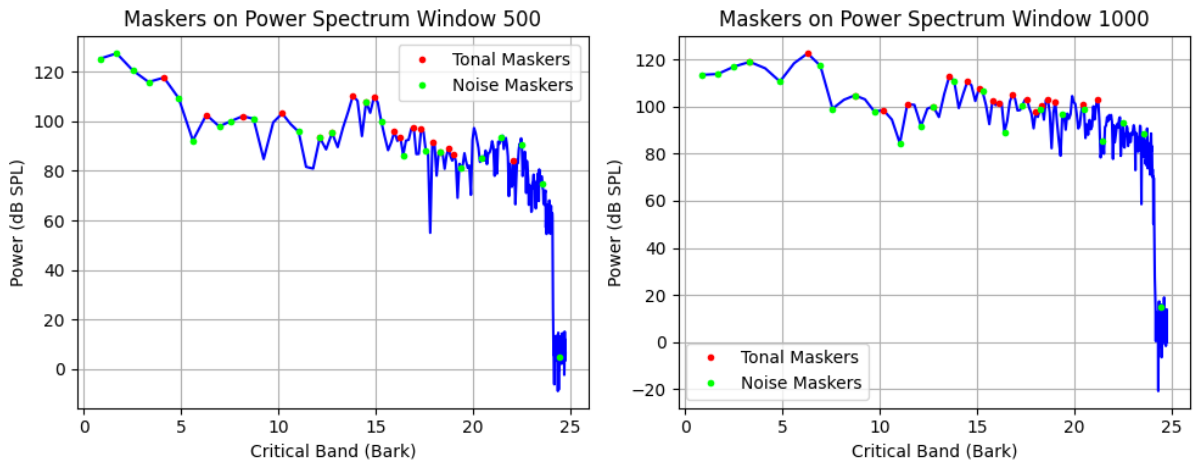
Η οποία επιστρέφει 0 αν η διακριτή συχνότητα k δεν αποτελεί τονική μάσκα, ενώ επιστρέφει 1 σε αντίθετη περίπτωση. Το εύρος περιοχών στο οποίο αναζητούμε τοπικά μέγιστα, αλλάζει ανάλογα με το που βρισκόμαστε στο πεδίο συχνότητας σύμφωνα με το Δ_k:

$$\Delta_k \in \begin{cases} 2, & 2 \le k < 63 \quad (0.17 - 5.5\text{kHz}) \\ [2,3] & 63 \le k < 127 \quad (5.5 - 11\text{kHz}) \\ [2,6] & 127 \le k < 250 \quad (11 - 20\text{kHz}) \end{cases}$$

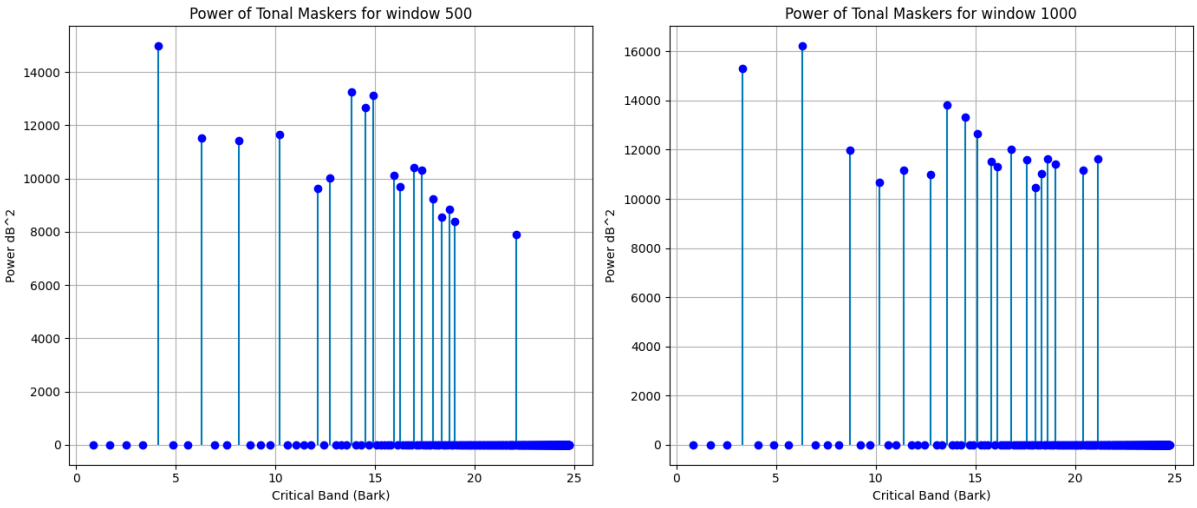
Στη συνέχεια, υπολογίζουμε την ισχύ κάθε μάσκας $P_{TM}(k)^2$ σύμφωνα με:

$$P_{TM}(k) = \begin{cases} 10\log_{10}(10^{0.1(P(k-1))} + 10^{0.1(P(k))} + 10^{0.1(P(k+1))})(\text{dB}), & \text{αν } S_T(k) = 1 \\ 0, & \text{αν } S_T(k) = 0 \end{cases}$$

Μας δίνεται επίσης ο προϋπολογισμένος πίνακας $P_{NM}(k)$, που περιέχει αντίστοιχα την ισχύ των μασκών θορύβου.
Για τα παράθυρα 500 και 1000 παίρνουμε λοιπόν τα παρακάτω:



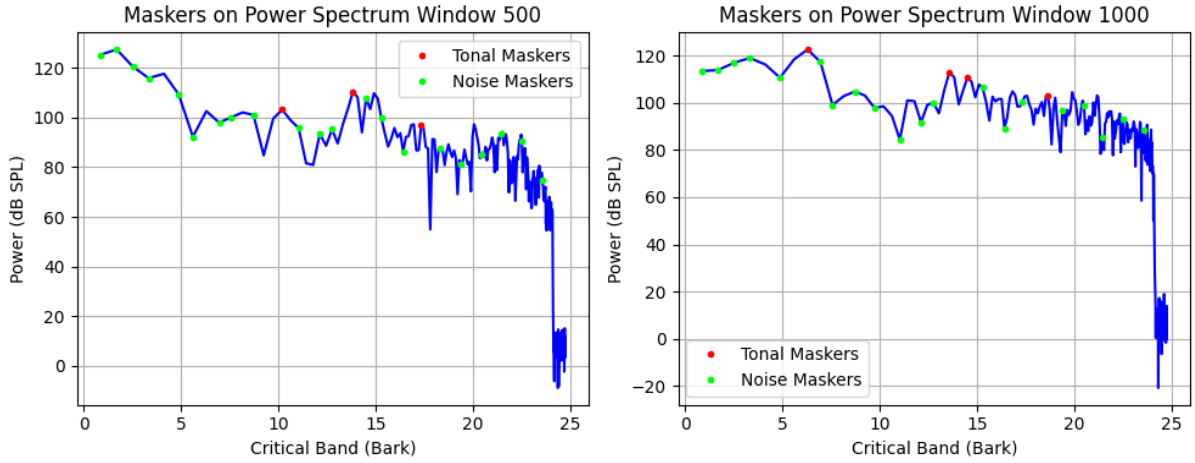
Εικόνα 3



Εικόνα 4

1.3 Μείωση και αναδιοργάνωση των μασκών

Μειώνοντας και αναδιοργανώντας τις μάσκες, μας δίνονται οι προϋπολογισμένοι πίνακας P_{TMc} και P_{NMc}. Συγκρίνοντας τα παρακάτω διαγράμματα (*Εικόνα 5*), με τα αντίστοιχα διαγράμματα πρωτού γίνει η μείωση *(Εικόνα 4)*, παρατηρούμε πως πράγματι ο αριθμός μασκών είναι μικρότερος.



Εικόνα 5

1.4 Υπολογισμός των δυο διαφορετικών κατωφλίων κάλυψης (Individual Masking Thresholds)

Υπολογίζουμε τα δύο διαφορετικά κατώφλια κάλυψης. Το κάθε κατώφλι αντιπροσωπεύει το ποσοστό κάλυψης στο σημείο i το οποίο προέρχεται από την μάσκα τόνου ή θορύβου στο σημείο j. Το δύο κατώφλια υπολογίζονται ως:

$$T_{TM}(i,j) = P_{TM}(j) - 0.275b(j) + SF(i,j) - 6.025(\text{dB SPL})$$

$$T_{NM}(i,j) = P_{NM}(j) - 0.175b(j) + SF(i,j) - 2.025(\text{dB SPL})$$

Όπου η συνάρτηση SF(i, j) υπολογίζει την έκταση της κάλυψης από το σημείο j στο οποίο βρίσκεται η μάσκα έως το σημείο i το οποίο υφίσταται κάλυψη και, για την περίπτωση των τονικών μασκών, μοντελοποιείται ως εξής:

$$SF(i,j) = \begin{cases} 17\Delta_b - 0.4P_{TM}(j) + 11, & -3 \le \Delta_b < -1 \\ (0.4P_{TM}(j) + 6)\Delta_b, & -1 \le \Delta_b < 0 \\ -17\Delta_b, & 0 \le \Delta_b < 1 \\ (0.15P_{TM}(j) - 17)\Delta_b - 0.15P_{TM}(j), & 1 \le \Delta_b < 8 \end{cases}$$

Θεωρούμε πως το κατώφλι T_{TM} ορίζεται σε ένα διάστημα γειτονιάς των 12-Bark της μάσκας στο σημείο j, δηλαδή στις θέσεις i : b(i) ∈[b(j) −3,b(j) + 8].

1.5 Υπολογισμός του συνολικού κατωφλίου κάλυψης (Global Masking Threshold)

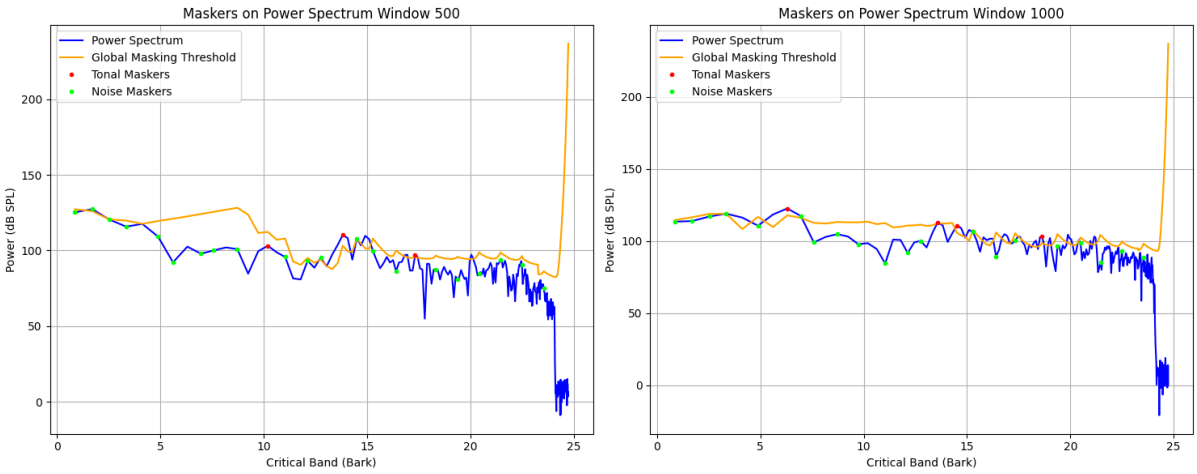
Τέλος, έχοντας υπολογίσει κάθε κατώφλι ξεχωριστά, υπολογίζουμε το συνολικό κατώφλι κάλυψης σε κάθε διακριτή συχνότητα ξεχωριστά σύμφωνα με τον τύπο:

$$T_g(i) = 10\log_{10}\left(10^{0.1T_q(i)} + \sum_{l=0}^{255} 10^{0.1T_{TM}(i,\ell)} + \sum_{m=0}^{255} 10^{0.1T_{NM}(i,m)}\right) \text{ dB SPL}$$

, όπου T_q το Absolute Threshold Hearing (ATH), το οποίο χαρακτηρίζει το ποσό της ενέργειας σε dB - Sound Pressure Level (dB SPL) που πρέπει να έχει ένας τόνος (π.χ. ημίτονο), συχνότητας f, ώστε να γίνει αντιληπτός σε περιβάλλον πλήρους ησυχίας. Ορίζεται από την παρακάτω σχέση:

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 \text{ (dB SPL)}$$

Συγκρίνουμε στα παρακάτω διαγράμματα (*Εικόνα 6*) το Global Masking Threshold, σε σχέση με το φάσμα ισχύος που υπολογίσαμε στο 1.1.



Εικόνα 6

Μέρος 2: Χρονο-Συχνοτική Ανάλυση με Συστοιχία Ζωνοπερατών Φίλτρων

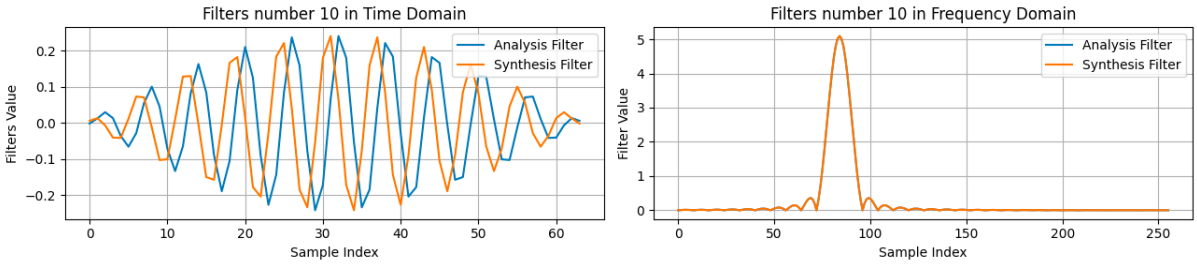
2.0 Συστοιχία Ζωνοπερατών Φίλτρων (Filterbank)

Όπως πριν παραθυροποιήσαμε το σήμα μουσικής στο πεδίο του χρόνου, έτσι και τώρα κάθε παράθυρο, το αναλύουμε στις κρίσιμες συχνότητες του χρησιμοποιώντας συστοιχίες φίλτρων (Filterbank). Συγκερκιμένα, ορίζουμε M = 32 φίλτρα ανάλυσης h_k και σύνθεσης g_k σύμφωνα με τις εξισώσεις:

$$h_k(n) = \sin\left[\left(n+\frac{1}{2}\right)\frac{\pi}{2M}\right]\sqrt{\frac{2}{M}}\cos\left[\frac{(2n+M+1)(2k+1)\pi}{4M}\right]$$

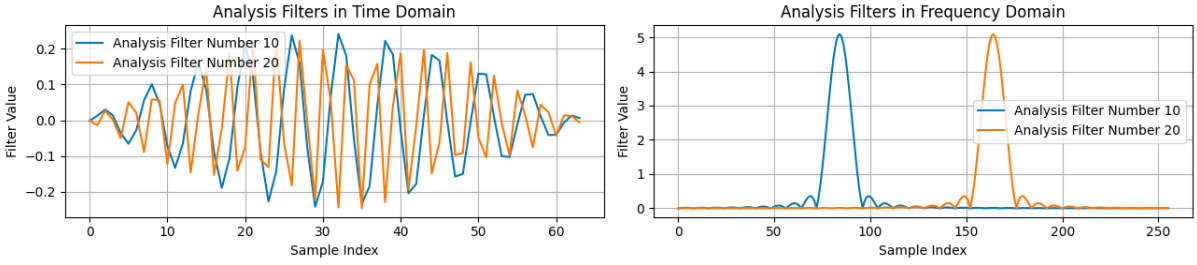
$$g_k(n) = h_k(2M-1-n)$$

Με μήκος L = 2M = 64.
Πρακτικά, το φίλτρο σύνθεσης g_k ισούται με το οριζόντια ανεστραμμένο φίλτρο h_k στο πεδίο του χρόνου. Αυτό σημαίνει ότι το συχνοτικό τους περιεχόμενο είναι ίδιο. Αυτό το συμπέρασμα φαίνεται στα παρακάτω διαγράμμα (*Εικόνα 7*).



Εικόνα 7

Επίσης συγκρίνουμε το πως διαφέρει το συχνοτικό περιεχόμενο σε μία συστοιχία φίλτρων, παίρνοντας για παράδειγμα το 10ο και το 20ο φίλτρο ανάλυσης:



Εικόνα 8

2.1 Ανάλυση με Συστοιχία Φίλτρων

Κάθε παράθυρο, το συνελίσουμε με τα 32 διαφορετικά φίλτρα ανάλυσης. Έτσι από ένα παράθυρο x_n προκύπτουν 32 νέα σήματα, τα οποία και υποδειγματοληπτούμε κατά τον παράγοντα M, λαμβάνοντας τελικά τα σήματα

$$y_k(n)$$

Όπου k = 0, 1.., 31

2.2 Κβαντοποίηση

Στη συνέχεια, κβαντοποιούμε το σήμα. Χρησιμοποιούμε δύο ήδη κβαντιστών:

Προσαρμοζόμενος ομοιόμορφος κβαντιστής 2^{B_k} επιπέδων:

Το B_k ισούται με τον αριθμό των bits κωδικοποίησης ανά δείγμα της ακολουθίας $y_k(n)$ στο τρέχον πλαίσιο ανάλυσης x_n του σήματος, ο οποίος και υπολογίζεται από τον τύπο:

$$B_k = \text{int}\left(\log_2 \left(\frac{R}{\min(T_g(i))} \right) - 1 \right)$$

Όπου T_g το Global Masking Threshold, όπως το υπολογίσαμε στο ερώτημα 1.5. Για την κβαντοποίηση του σήματος χρησιμοποιούμε τη συνάρτηση **digitize()** της **numpy**.

Μη - Προσαρμοζόμενος κβαντιστης 2^{B_k} επιπέδων:

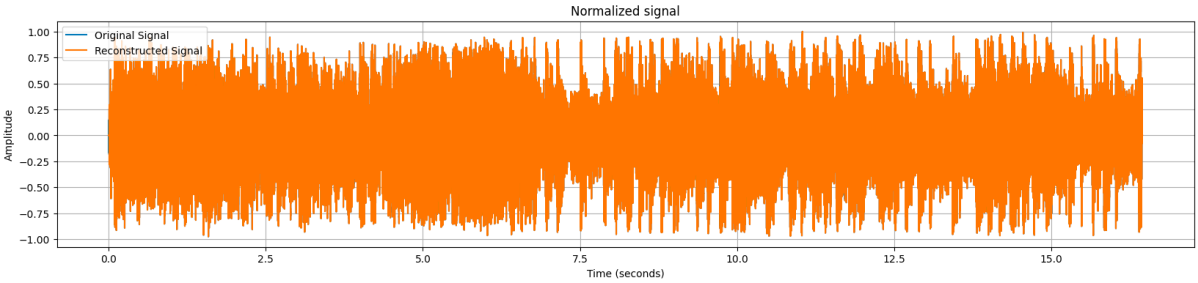
Χρησιμοποιούμε σταθερό αριθμό bits κωδικοποίησης B_k = 8 bits.

2.3 Σύνθεση

Τέλος, κάθε κβαντοποιημένο y_k, συνελίσσεται με το αντίστοιχο φίλτρο σύνθεσης g_k και στην συνέχεια υπερδειγματοληπτείται. Για τη διαδικασία της υπερδειγματοληψίας, δημιουργήσαμε μία δική μας συνάρτηση **upsample()**. Τέλος χρησιμοποιώντας την μέθοδο OverLap-Add ανακατασκευάζουμε το σήμα μουσικής.

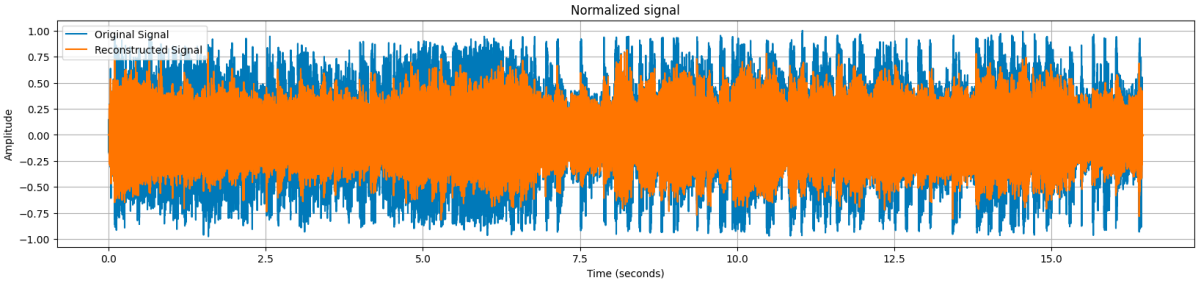
Μέρος 3: Συμπεράσματα

Σχεδιάζουμε τα ανακατασκευασμένου σήμα, από τον προσαρμοζόμενο κβανιστή, στο ίδιο διάγραμμα με το αρχικό σήμα μουσικής (*Εικόνα 9*):



Εικόνα 9

Παρατηρούμε πως τα δύο σήματα είναι αδιαχώριστα. Αντίθετα, κάνοντας το ίδιο για τον μη- προσαρμοζόμενο κβαντιστή παίρνουμε:



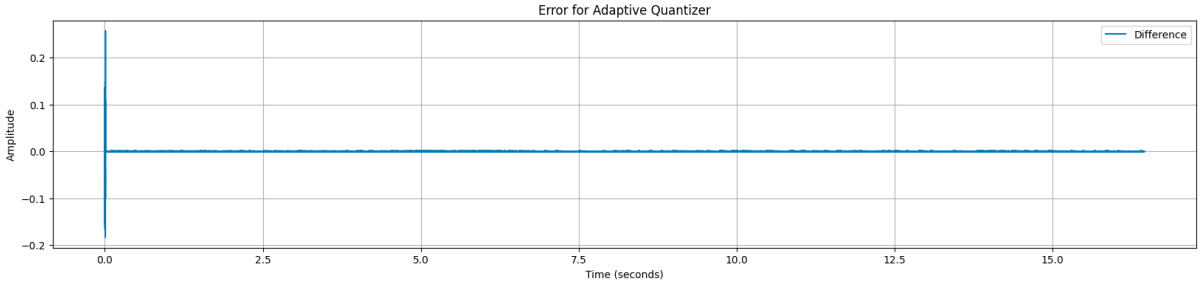
Εικόνα 10

Ήδη, και με το μάτι, φαίνεται πόσο καλύτερη δουλειά έχει γίνει με τον προσαρμοζόμενο κβανιστή. Ίδια συμπεράσματα βγάζουμε ακούγοντας επιπλέον, τα δύο ανακασκευασμένα σήματα. Ο προσαρμοζόμενος παράγει σήμα που ακούγεται ακριβώς όπως το αρχικό. Αντίθετα ο μη-προσαρμοζόμενος, δημιουργεί σήμα που έχει εμφανή θόρυβο. Το αρχικό σήμα μουσικής εξακολουθεί να ακούγεται με ευκολία αλλά σίγουρα έχει πέσει η ποιότητα ήχου.
Σχετικά με το κατά πόσο οι δύο κβαντιστές καταφέρνουν να συμπιέσουν το αρχικό σήμα, λαμβάνουμε υπόψιν μετρικές όπως Mean Square Error (mse) και Compression Rate:

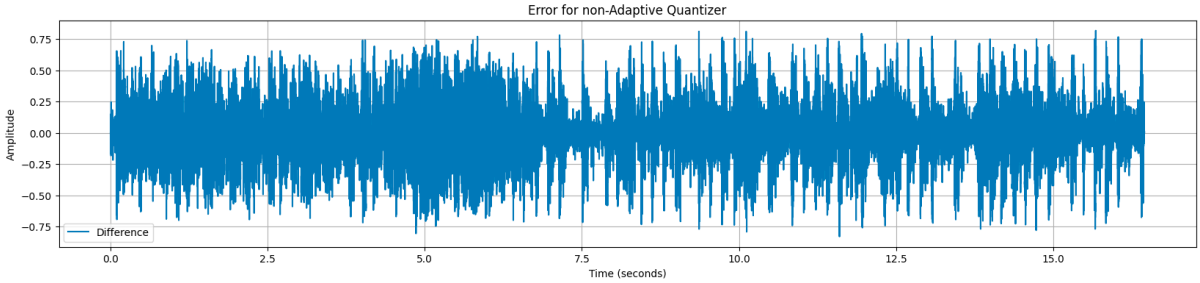
	M.S.E	Compression Rate	Total Number of Bits
Adaptive Quantizer	3.821e-06	0.305	8,060,580
Non-Adaptive 8bit Quantizer	25,000e-06	0.437	6,529,536

Παρατηρούμε ότι, ο μη-προσαρμοζόμενος, καταφέρνει να κάνει καλύτερη συμπίεση, όσον αφορά τον χώρο, όμως με σημαντικό κόστος στην ποιότητα ήχου. Το mse μάλιστα δείχνει πόσο σημαντική είναι η απόκλιση του από το πραγματικό σήμα.

Τέλος, παραθέτουμε τα διαγράμματα που απεικονίζουν το λάθος (διαφορά αρχικού και ανακατασκευασμένου) σε κάθε περίπτωση:



Εικόνα 11



Εικόνα 12