Ψηφιακή Επεξεργασία Σημάτων
3η Εργαστηριακή Άσκηση

Θέμα: Κωδικοποίηση σημάτων Μουσικής βάσει ψυχοακουστικού μοντέλου (Perceptual Audio Coding)

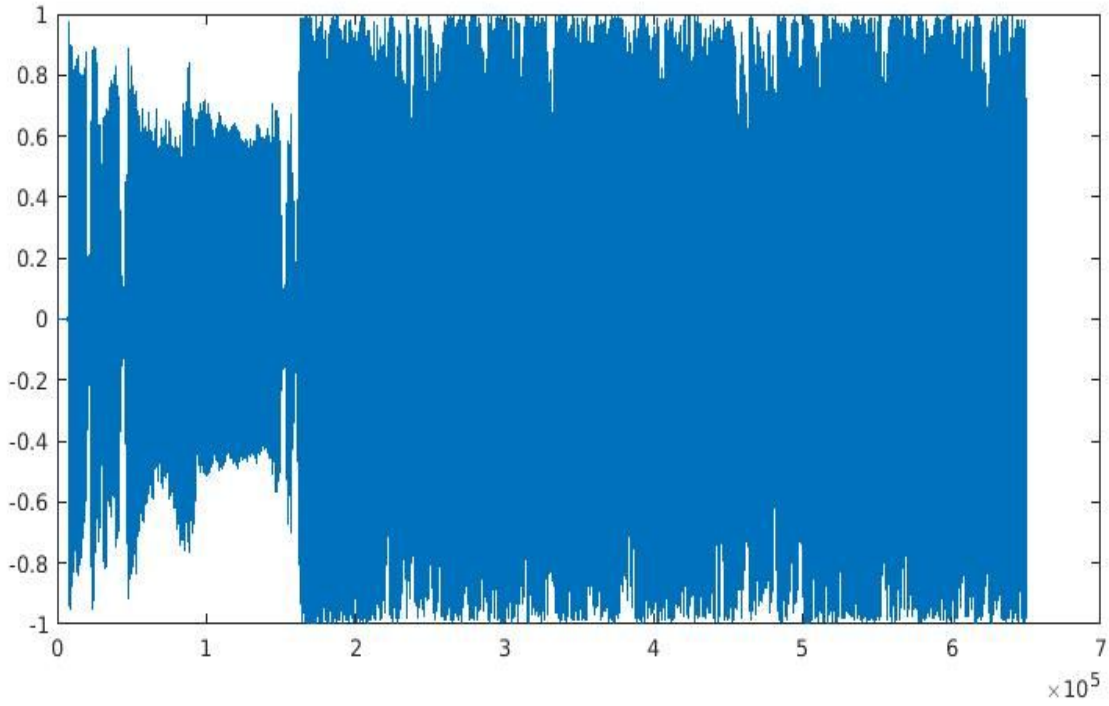
Μαρία Παρέλλη *Γεώργιος-Ορέστης Χαρδούβελης*
03115155 *03115100*
6ο Εξάμηνο *6ο Εξάμηνο*

Μέρος 1: Ψυχοακουστικό Μοντέλο 1

Στην παρούσα άσκηση καλούμαστε να κωδικοποιήσουμε ένα σήμα μουσικής με στόχο την βελτιστοποίηση της επεξεργασίας και μείωσης των αριθμών bits που απαιτούνται για την κωδικοποίηση και αποθήκευση του σήματος (συμπίεση), διατηρώντας παράλληλα την ποιότητα του.Ένας τρόπος να επιτευχθεί αυτό είναι μέσω της εκμετάλλευσης των φυσικών ελαττωμάτων του ανθρώπινου αυτιού-δέκτη καθώς και το ψυχοακουστικό μοντέλο.

1.0

Καταρχάς, το μουσικό μας σήμα βρίσκεται σε δύο κανάλια, οπότε τα συνδυάζουμε σε ένα υπολογίζοντας τον μέσο όρο των αντίστοιχων bits από κάθε κανάλι. Στην συνέχεια, για να γίνει κανονικοποίηση στο διάστημα [-1,1], βρίσκουμε την μέγιστη τιμή (σε απόλυτο) και διαιρούμε κάθε τιμή του σήματος μας με αυτή. Το σήμα, συναρτήσει των δειγμάτων του απεικονίζεται ως εξής:



1.1

Αρχικά ορίζουμε την κλίμακα Bark ως εξής:

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

όπου f ο πίνακας συχνοτήτων μας σε Hz.

Κατά αυτόν τον τρόπο μετατρέπουμε τις συχνότητες μας από Hz σε Bark. Ο λόγος που γίνεται η παραπάνω διαδικασία είναι για να οριστούν τα **critical bands**, τα οποία αντιστοιχούν στις περιοχές στις οποίες συντονίζονται οι νευροδέκτες του ακουστικού φλοιού. Συγκεκριμένα, οι 25 πρώτες κρίσιμες συχνοτικές περιοχές μοντελοποιούνται με την ψυχοακουστική κλίμακα συχνοτήτων Bark η οποία έχει πεδίο τιμών στο διάστημα [1, 25].

Ύστερα ορίζουμε και το κατώφλι ακοής (**Absolute Threshold of Hearing**) που χαρακτηρίζει το ποσό της ενέργειας σε dB - Sound Pressure Level (dB SPL) που πρέπει να έχει ένας τόνος συχνότητας f ώστε να γίνει αντιληπτός σε περιβάλλον πλήρους ησυχίας.

Ο τύπος έχει ως εξής:

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

Σε εφαρμογές συμπίεσης, το Tq(f) θα μπορούσε να ερμηνευθεί ως το μέγιστο επιτρεπτό ποσό ενέργειας για την κωδικοποίηση των παραμορφώσεων που εισάγονται στο πεδίο της συχνότητας.

Στην συνέχεια, παραθυρώνουμε το σήμα μας με παράθυρο Hanning και μήκος παραθύρου L=512. Έτσι προκύπτουν 1271 παράθυρα και εκτελούμε όλα τα παρακάτω βήματα για κάθε παράθυρο ξεχωριστά.

1.1

Αφού ορίσαμε την κλίμακα Bark, υπολογίζουμε το φασματικό περιεχόμενο του σήματος, το οποίο σε φυσικά μεγέθη δηλώνει την πίεση που υφίσταται το ανθρώπινο αυτί και μετατρέπουμε τη συχνότητα από Hz σε Bark.

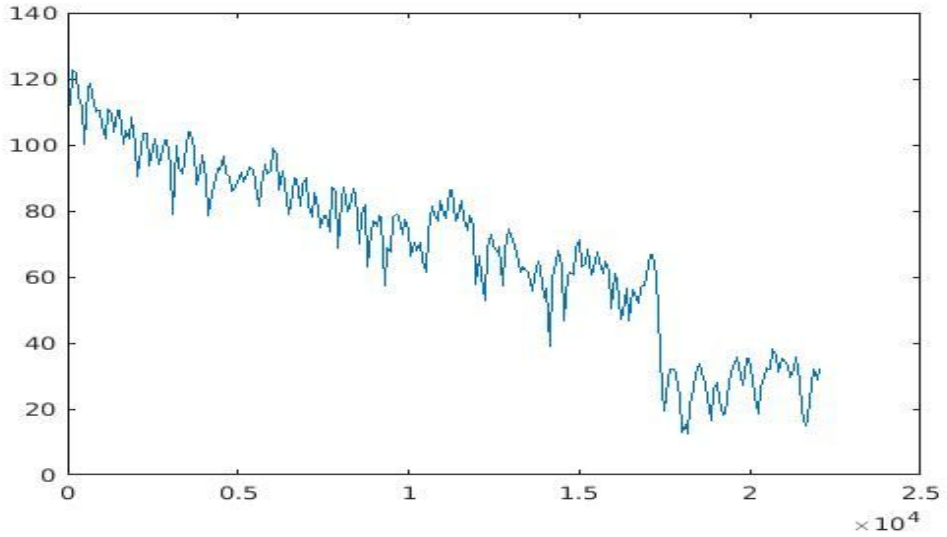
Ακολούθως υπολογίζουμε το N-σημείων φάσμα ισχύος P(k) του σήματος όπου N = 512 δείγματα όπως έχει καθιερωθεί στο πρότυπο MPEG Layer-1.

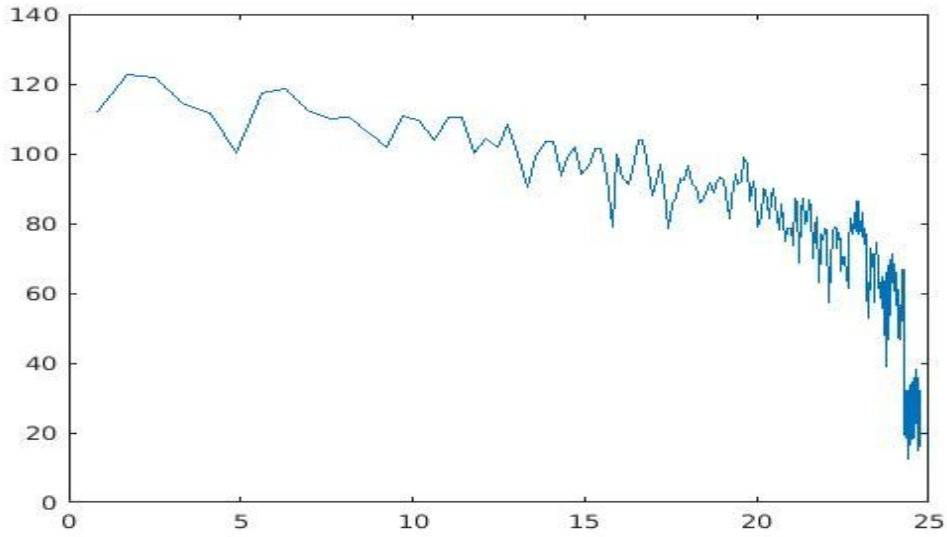
$$P(k) = PN + 10\log_{10}\left|\sum_{n=0}^{N-1} w(n)x(n)e^{-j\frac{2\pi kn}{N}}\right|^2, 0 \le k \le \frac{N}{2}.$$

με PN=90.302 dB και στο w(n) χρησιμοποιούμε το παράθυρο Hanning που ορίζεται ως εξής:

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

Παρακάτω παρατίθενται οι γραφικές παραστάσεις της ισχύος του σήματος,τόσο σε κλίμακα Hertz όσο σε κλίμακα Bark.





1.2: Εντοπισμός μασκών τόνων και θορύβου

Αφού υπολογιστεί το φάσμα ισχύος P(k), στη συνέχεια εντοπίζουμε ανά critical band τοπικά μέγιστα (μάσκες) τα οποία είναι μεγαλύτερα από τις γειτονικές τους συχνότητες τουλάχιστον κατα 7 dB.

Εδώ πρέπει να δοθούν οι εξής ορισμοί:

Tone Maskers: Τόνοι που βρίσκονται στο μέσο μιας κρίσιμης μπάντας και καλύπτουν το θόρυβο ενός τόνου κατωτέρας σημασίας για την πληροφορία του σήματος.
Noise Maskers: Τόνοι όμοιοι με τους tone maskers όπου εδώ ο θόρυβος λειτουργεί σαν μάσκα άλλων τόνων, αντιστέφοντας την παραπάνω διαδικάσια.

Η ισχύς της τονικής μάσκας στη θέση k ορίζεται ως εξής:

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

Η συνάρτηση $S_T(k)$ είναι μια boolean συνάρτηση (έχει σύνολο τιμών [0,1]) και πρακτικά ορίζει αν υπάρχει η όχι τονική μάσκα στο σημείο k, και ορίζεται ως εξής:

$$S_T = \begin{cases} 0, & \text{αν } k \notin [3, 250] \\ P(k) > P(k \pm 1) \bigwedge P(k) > P(k \pm \Delta_k) + 7\text{dB}), & \text{αν } k \in [3, 250] \end{cases}$$

Με το Δ να ορίζεται ως εξής:

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

Για την εύρεση των μασκών θορύβου (noise makers) χρησιμοποιούμε την έτοιμη συνάρτηση findNoiseMaskers που παίρνει ως ορίσματα το φάσμα ισχύος P του σήματος, την ισχύ τονικής μάσκας Ptm, και η κλίμακα συχνοτήτων Bark b.

Στο παράδειγμα του 450ου παραθύρου που χρησιμοποιούμε, οι αρχικές μάσκες τόνου και θορύβου που βρίσκουμε εντοπίζονται στις θέσεις:

- Ptm:

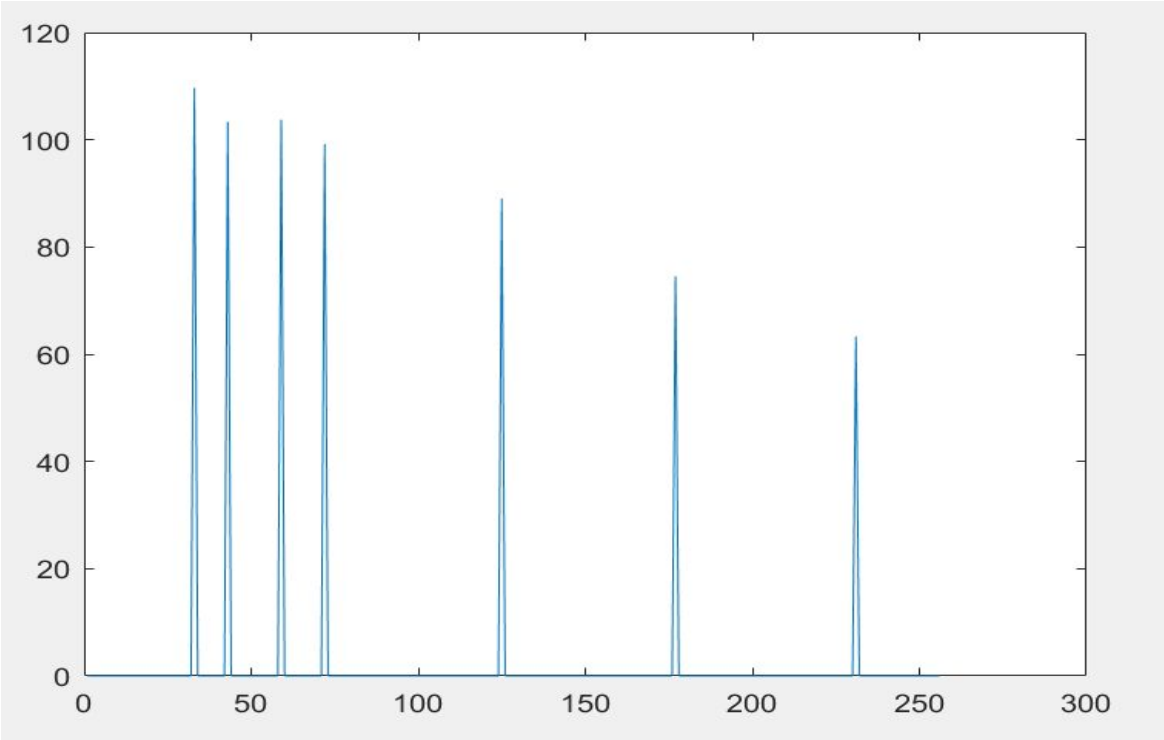
```
>> ask2_2
    33    43    59    72   125   177   231
```

- Pnm

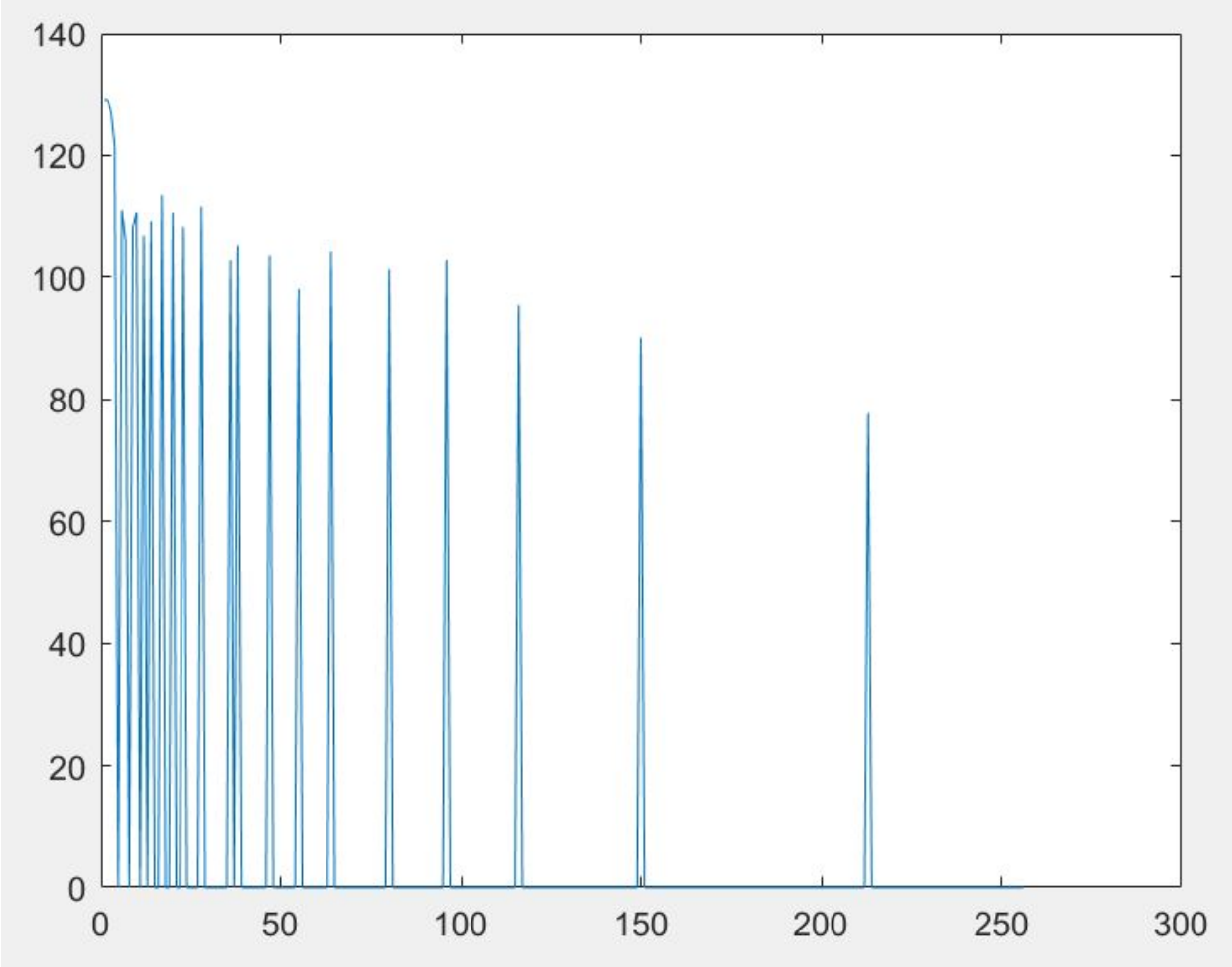
```
>> ask2_2
  Columns 1 through 20

     1     2     3     4     6     7     9    10    12    14    17    20    23    28    36    38    47    55    64

  Columns 21 through 24

    96   116   150   213
```

Τα διαγράμματα Ptm καθώς και Ptn απεικονίζονται παρακάτω:

- Ptm



- Pnm



1.3

Στο βήμα αυτό μειώνουμε και αναδιοργανώνουμε τις μάσκες του προηγούμενου παραδείγματος, χρησιμοποιώντας τη συνάρτηση:

[PTM,PNM] =checkMaskers(PTM,PNM,Tq,b)

Πρακτικά, με τη συνάρτηση αυτή απορρίπτονται μάσκες οι οποίες βρίσκονται κάτω από το κατώφλι απόλυτης ακοής, καθώς και μάσκες ασθενέστερες από άλλες κοντινές τους.
Οι μάσκες που απομένουν μετά την κλήση της συνάρτησης βρίσκονται στα εξής σημεία:

- Ptm

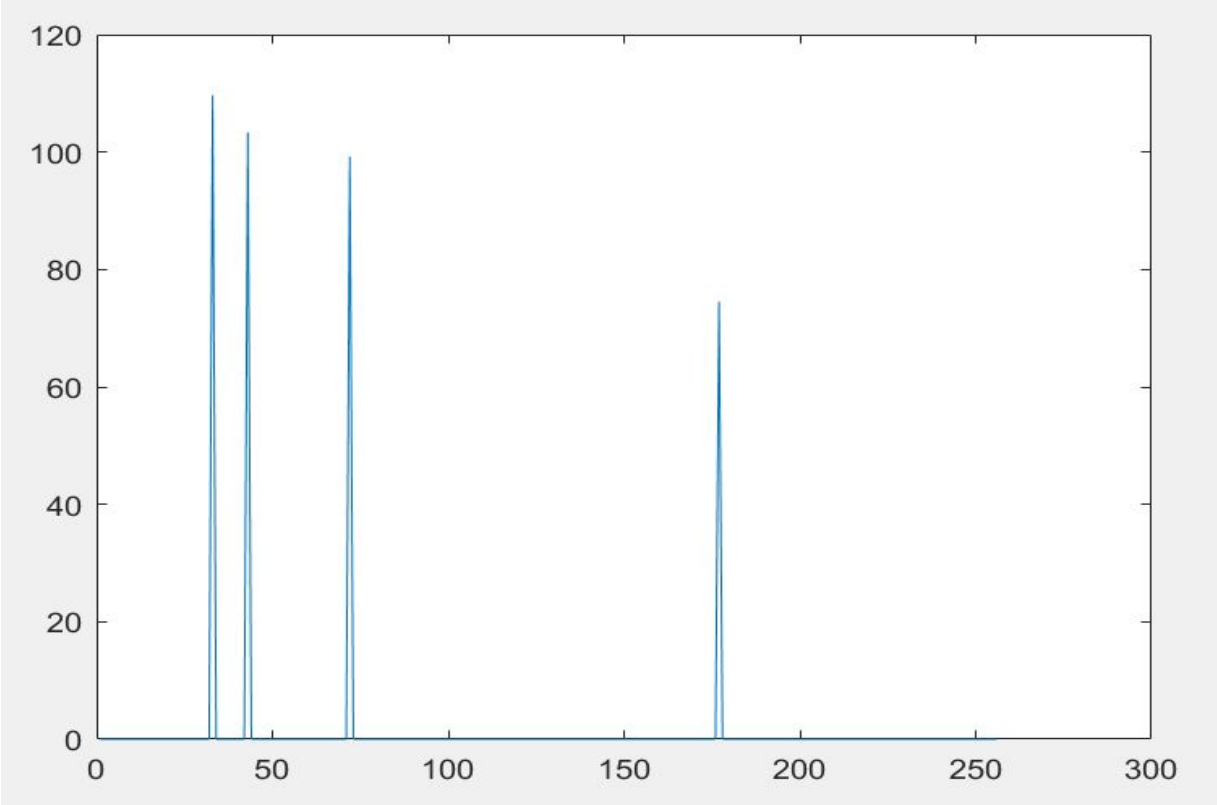
```
>> ask2_2
    33    43    72   177
```

- Pnm

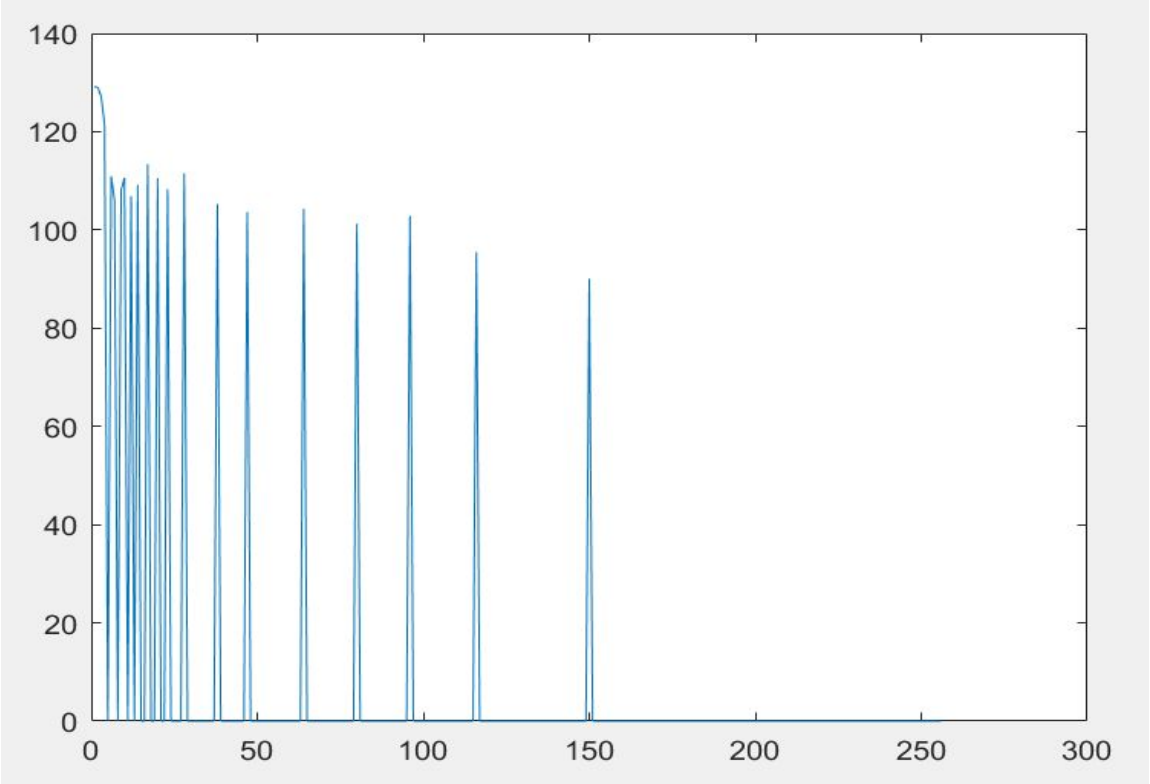
```
Columns 1 through 19

   1     2     3     4     6     7     9    10    12    14    17    20    23    28    38    47    64    80    96

Columns 20 through 21

  116   150
```

Ταυτόχρονα, οι γραφικές τους παραστάσεις έχουν ως εξής:

- Ptm



- Pnm



1.4

Μετά την μείωση του αριθμού των μασκών , υπολογίζουμε τα δύο διαφορετικά κατώφλια κάλυψης. Το κάθε κατώφλι αντιπροσωπεύει το ποσοστό κάλυψης σε ένα σημείο i το οποίο προέρχεται από την μάσκα τόνου ή θορύβου σε ένα σημείο j.

Τα κατώφλια υπολογίζονται ως εξής:

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

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

Όπου η SF υπολογίζει την έκταση της κάλυψης από το σημείο j στο οποίο βρίσκεται η μάσκα έως το σημείο i το οποίο υφίσταται κάλυψη και μοντελοποιείται ως εξής :

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

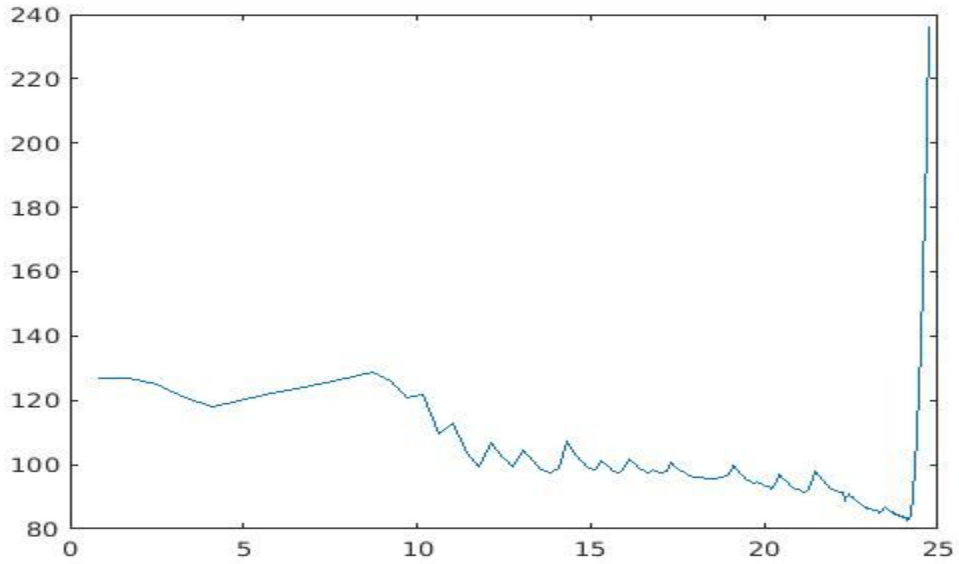
1.5

Τέλος, ορίζουμε το global κατώφλι κάθε δείγματος για κάθε πλαίσιο,αθροιστικά με τον εξής τρόπο:

$$T_g(i) = 10\log_{10}\left(10^{0.1T_q(i)} + \sum_{l=1}^{L} 10^{0.1T_{TM}(i,\ell)} + \sum_{m=1}^{M} 10^{0.1T_{NM}(i,m)}\right) \text{ dB SPL}$$

Για τον υπολογισμό των επιμέρους κατωφλίων κάλυψης $T_{TM}(i,m)$ και $T_{NM}(i,m)$ για κάθε critical band , που αντιστοιχεί στις διαφορετικές μάσκες, αθροίζονται τα επιμέρους κατώφλια.

Για το πλαίσιο 450 η εικόνα του Tg έχει ώς εξής :



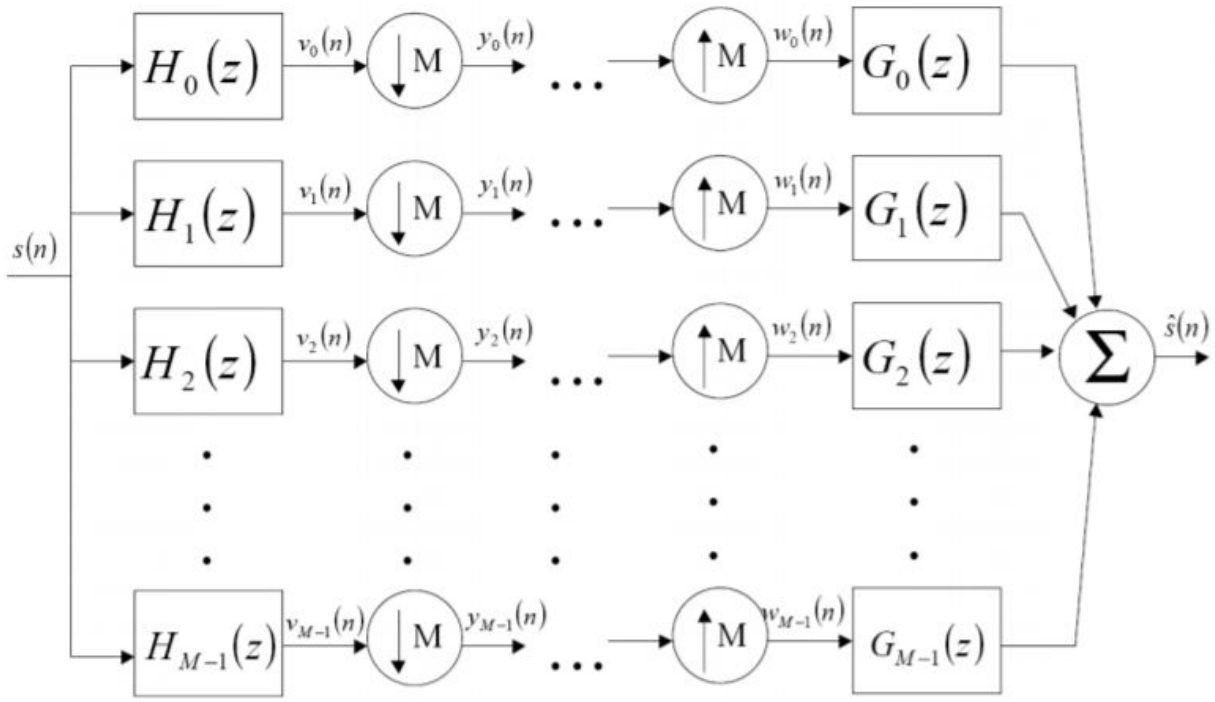
Παρατηρούμε μια -σε μεγάλο βαθμό- σταθερή τιμή, η οποία αυξάνεται απότομα στο τέλος της κλίμακας bark, με αποτέλεσμα να μπορούμε να αποφύγουμε την κωδικοποίηση των τελευταίων συχνοτήτων του παραθύρου.

Να τονιστεί ότι η συνάρτηση που υπολογίζει το Tg επιστρέφει έναν δισδιάστατο πίνακα, στον οποίον το πλήθος των γραμμών αντιστοιχεί στα πλαίσια (στα οποία έχει διαιρεθεί το σήμα) και η κάθε γραμμή περιέχει το global masking threshold που αντιστοιχεί στο πλαίσιο αυτό.

Μέρος 2. Χρονο-Συχνοτική Ανάλυση με Συστοιχία Ζωνοπερατών Φίλτρων

Σε αυτό το μέρος φιλτράρουμε τα παράθυρα x(n) του σήματος με συστοιχία φίλτρων ανάλυσης hk(n) και σύνθεσης gk(n). Έτσι υλοποιείται η διαδικασία του παρακάτω σχήματος, η οποία παίρνει σαν είσοδο το κάθε πλαίσιο ανάλυσης x(n), τη συστοιχία φίλτρων και το συνολικό κατώφλι

κάλυψης που υπολογίστηκε στο Μέρος 1. Η έξοδος είναι το ανακατασκευασμένο σήμα $\widehat{x}$ (n).



Δεδομένου του κατωφλιού κάλυψης που υπολογίστηκε στο προηγούμενο μέρος, προσδιορίζουμε τις περιπτώσεις στις οποίες η κβάντιση του ήχου είναι περιττή, είτε λόγω της ύπαρξης υψηλών εντάσεως τόνων στην ίδια συχνοτική περιοχή, είτε λόγω της ιδιομορφίας της ανθρώπινης ακοής να αντιλαμβάνεται ορισμένες συχνότητες περισσότερο από άλλες.

2.0

Στο βήμα αυτό ορίζουμε τις συστοιχίες των ζωνοπερατών φίλτρων που φαίνονται στο παραπάνω σχήμα. Ορίζονται 32 φίλτρα ανάλυσης και 32 φίλτρα σύνθεσης με μήκος 64 το καθένα, τα οποία ορίζονται αντίστοιχα ως εξής:

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

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

2.1

Από εδώ και στο εξής λειτουργούμε σε κάθε πλαίσιο του κανονικοποιημένου σήματος μας ξεχωριστά επαναλαμβάνοντας την ίδια διαδικασία.
Αρχικά, παίρνουμε την συνέλιξη του πλαισίου με την συστοιχία των φίλτρων ανάλυσης hk μέσω της συνάρτησης **conv**.

$$v_k(n) = h_k(n) * x(n) = \sum_{m=0}^{L-1} x(n - m)h_k(m),\ k = 0, 1, \ldots, M - 1$$

Έτσι προκύπτει το φιλτραρισμένο σήμα yk(n), στο οποίο εφαρμόζουμε αποδεκατισμό (decimation) κατά παράγοντα M, μέσω της εντολής **downsample**. Έχουμε λοιπόν:

$$y_k(n) = v_k(Mn)$$

Ο παράγοντας αποδεκατισμού επιλέγεται έτσι ώστε να έχουμε τον μέγιστο αποδεκατισμό δίχως να υπάρχουν επικαλύψεις. Δεδομένου της μη ιδανικότητας των ζωνοπερατών φίλτρων υπάρχουν μικρές επικαλύψεις, οι οποίες θεωρούνται αμελητέες.

2.2

Σε αυτό το βήμα υπολογίζουμε το κβαντισμένο σήμα μας, συνεχίζοντας να εργαζόμαστε σε κάθε πλαίσιο χωριστά.

Τα bit που θα χρησιμοποιηθούν υπολογίζονται από την Bk και έτσι θα έχουμε κβαντιστή με 2^(Bk) επίπεδα. Το μέγιστο ανεκτό σφάλμα συνδέεται με το συνολικό κατώφλι κάλυψης Tg(i) του ψυχοακουστικού μοντέλου όπως προέκυψε από το Μέρος 1 και η σχέση υπολογισμού του είναι :

$$B_k = \left\lceil \log_2 \left(\frac{R}{min(T_g(i))} - 1 \right) \right\rceil ,$$

όπου το R συμβολίζει το πλήθος βαθμίδων έντασης του αρχικού σήματος. Εφόσον το αρχικό σήμα έχει κωδικοποιηθεί με PCM χρησιμοποιώντας 16 bits ανά δείγμα, έχουμε R=2^16.

Το Tg(i) ορίζεται στο διάστημα ορισμού του κάθε φίλτρου ανάλυσης, δηλαδή i : f(i) ∈ [fk – Fsπ/2M] , [fk + Fsπ/2M], με fk να είναι η κεντρική συχνότητα του κάθε φίλτρου, fk=(2*k-1)*Fs * pi/(2*M).

Το βήμα της κβάντισης προσαρμόζεται στο κάθε πλαίσιο ανάλυσης και είναι ίσο με D=range/2^Bk, με range να είναι η διαφορά της μέγιστης από την ελάχιστη τιμή στο αντίστοιχο πλαίσιο του σήματος μας.
Βεβαίως, όλα τα παραπάνω υπολογίζονται για κάθε ακολουθία yk που έχει προκύψει από τα 32 διαφορετικά φίλτρα ανάλυσης.

Στη συνέχεια, πραγματοποιείται η κβάντιση μέσω της συνάρτησης quantiz, που παίρνει σαν ορίσματα

1. το σήμα που έχει προκύψει από το προηγούμενο βήμα
2. τον πίνακα partition που ξεκινάει από την ελάχτιστη τιμή του πλαισίου μας συν το αντίστοιχο βήμα μέχρι την μέγιστη, με βήμα D(k)
3. τον πίνακα codebook (που προσδιορίζει τις κβαντισμένες τιμές που θα παίρνουν οι τιμές του σήματος από τα αντίστοιχα διαστήματα που ορίζονται από τον partition) .Ο πίνακας ξεκινάει από την ελάχιστη τιμή του πλαισίου μας μέχρι την μέγιστη με βήμα D(k).

Τέλος,οι τιμές των κβαντισμένων πλασίων για κάθε ακολουθία yk αποθηκεύονται στον πίνακα y_q.

Σημειώνουμε σε αυτό το σημείο ότι η κβάντιση που θα πραγματοποιήσουμε είναι ομοιόμορφη, δηλαδή το βήμα μεταξύ δύο διαδοχικών σταθμών είναι σταθερό. Το γεγονός αυτό περιορίζει την απόδοση, μιας και η κατανομή της έντασης του σήματος δεν είναι ομοιόμορφη.

Στη συνέχεια επαναλαμβάνουμε την κβάντιση αλλά αυτή την φορά με έναν μη-προσαρμοζόμενο κβαντιστή, με σταθερό αριθμό bit του κβαντιστή, όπου Bk = 8 και σταθερό δήμα κβαντισμού Δ, το οποίο καθορίζεται από ένα υποτιθέμενο σταθερό πεδίο τιμών του σήματος [−1, 1].

2.3
Σε αυτό το σημείο οι κβαντισμένες ακολουθίες yk υπερδειγματοληπτούνται με παράγοντα M ως εξής:

$$w_k(n) = \begin{cases} \hat{y}_k(n/M), & n = 0, M, 2M, 3M, \dots \\ 0, & \textbf{αλλιώς.} \end{cases}$$

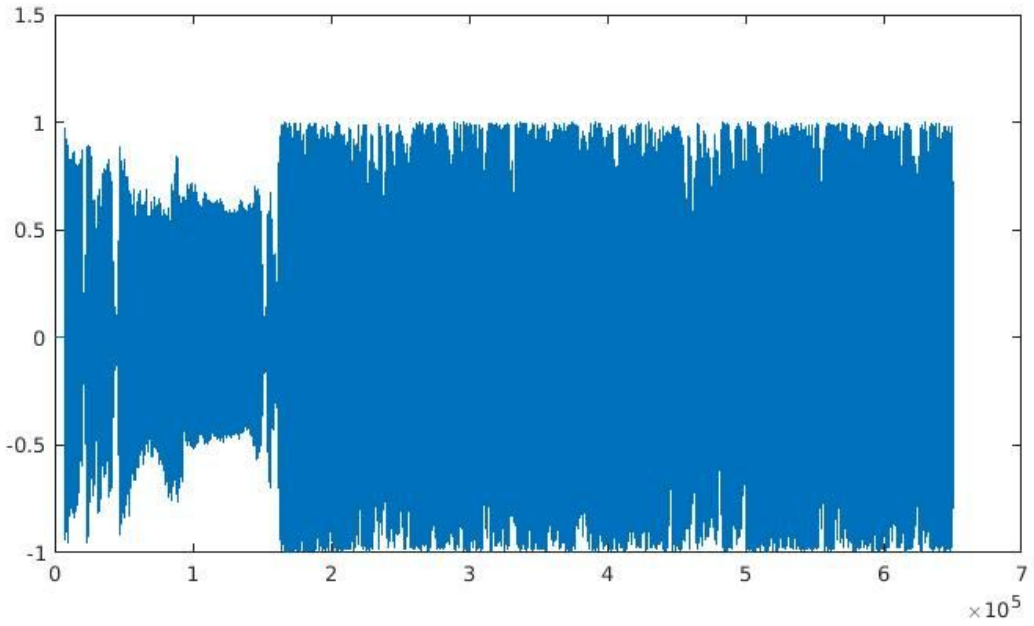
Στον κώδικα μας αυτό πραγματοποιείται με την συνάρτηση upsample.

Στη συνέχεια τα wk συνελίσονται με τα φίλτρα εξίσωσης gk που έχουν υπολογιστεί παραπάνω. Το αποτέλεσμα για κάθε διαφορετικό πλαίσιο αποθηκεύεται στον πίνακα s_w, όπου έχουμε και το τελικό μας σήμα σε πλαίσια.
Παρατηρούμε πως το μήκος κάθε πλαισίου από 512 έχει γίνει 639. Η αλλαγή αυτή οφείλεται τόσο στις συνελίξεις που πραγματοποιήθηκαν,

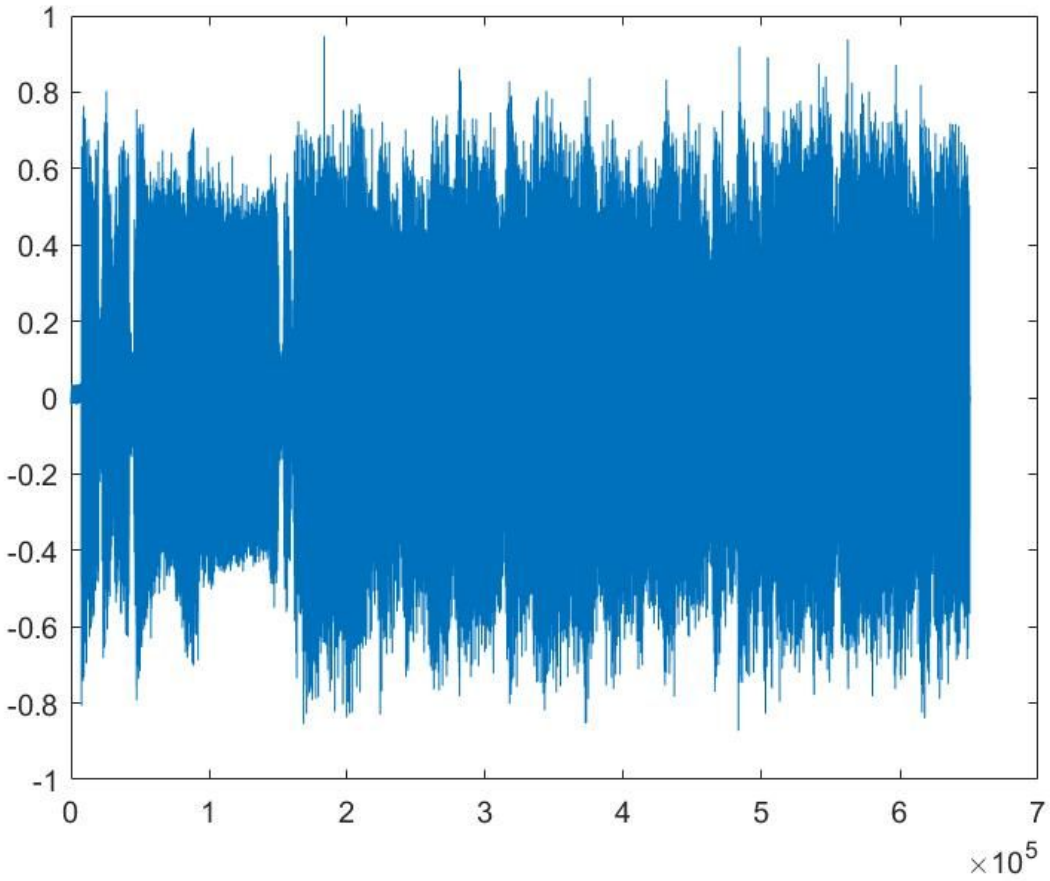
όσο και στις διαδικασίες downsample και upsample. Επομένως, στην δημιουργία του τελικού διανύσματος που αποτελεί και το επεξεργασμένο σήμα μας, η επιπλέον πληροφορία προστίθεται στη αρχή κάθε επόμενου παραθύρου σύμφωνα με την τεχνική overlap-add.
Τέλος, είναι σημαντικό να αναφέρουμε πως τα φίλτρα σύνθεσης εισάγουν καθυστέρηση στο συμπιεσμένο σήμα, η οποία, ειδικά στον υπολογισμό του μέσου τετραγωνικού λάθους πρέπει να αφαιρεθεί. Συγκεκριμένα, παρατηρείται μετατόπιση 2*M=64 bits.

Αποτελέσματα Σύγκρισης Κβαντιστών και Γραφικές Παραστάσεις

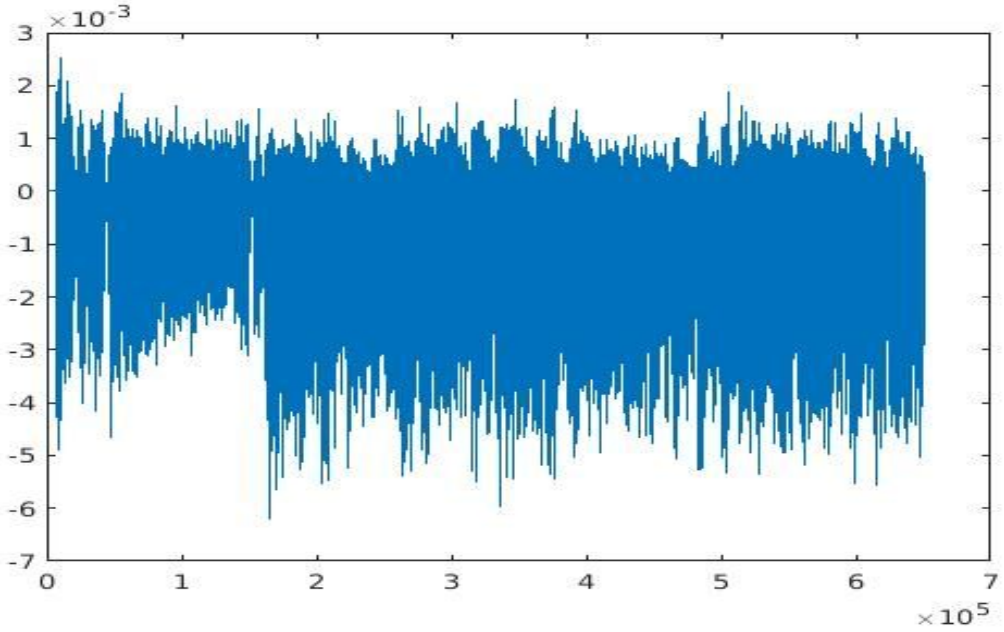
Plot του τελικού σήματος με προσαρμοσμένο κβαντιστή:



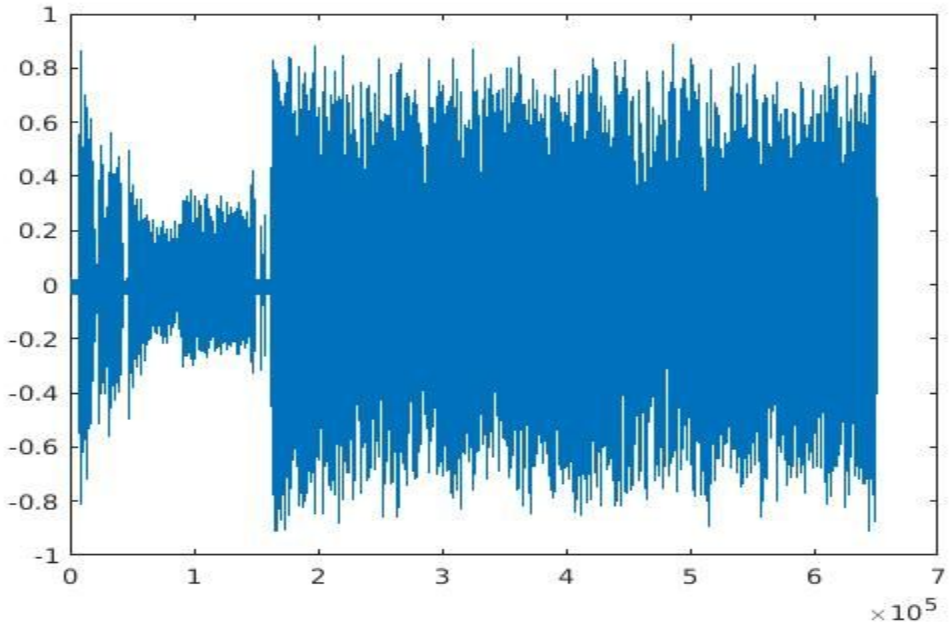
Plot του τελικού σήματος με σταθερό κβαντιστή:



Plot λάθους με προσαρμοσμένο κβαντιστή:



Plot λάθους με σταθερό κβαντιστή:



Υπολογισμός ποσοστού κβάντισης σε προσαρμοσμένο κβαντιστή

Για να προσδιοριστεί το παραπάνω μέγεθος πρέπει να υπολογιστεί ο μέσος όρος των bits, που χρησιμοποιήθηκαν για την κωδικοποίηση κάθε δείγματος του σήματος. Αυτός προκύπτει:

mean=9.8971

Τελικά, το ποσοστό κβάντισης στον προσαρμοσμένο κβαντιστή είναι:61.86%

Αντίθετα, στον σταθερό κβαντιστή που έχουμε επιλέξει χρησιμοποιούμε 8 bits κάθε φορά και το ποσοστό κβάντισης είναι 50%.

Υπολογισμός μέσου τετραγωνικού λάθους (MSE):

Εδώ υπολογίζουμε το μέσο τετραγωνικό λάθος του σήματος που προκύπτει σε σχέση με το αρχικό σήμα (σε ένα κανάλι και κανονικοποιήμενο). Ο υπολογισμός γίνεται μέσω της συνάρτησης immse.

Σημειώνουμε ξανά πως έχουμε λάβει υπόψη την μετατόπιση του τελικού σήματος κατά 64 bits λόγω των φίλτρων που χρησιμοποιήθηκαν.
Αφού τα υπολογίσουμε έχουμε:

- Χρησιμοποιώντας προσαρμοσμένη κβάντιση:
  MSE = 8.5761*10^(-7) = 8.58*10^(-5) %
- Χρησιμοποιώντας ΄σταθερής κβάντιση:
  MSE = 0.0545 = 5.45%

Όπως ήταν αναμενόμενο, στην περίπτωση του προσαρμοσμένου κβαντιστή έχουμε πολύ μικρότερο μέσο τετραγωνικό σφάλμα και εμφανέστερα καλύτερη ποιότητα σήματος. Αυτό ήταν αναμενόμενο εφόσον ο σταθερός κβαντιστης δεν προσαρμόζεται στα χαρακτηριστικά του κάθε παραθύρου σήματος.