Ψηφιακή Επεξεργασία Σήματος

2η Εργαστηριακή Άσκηση

Διαμάντη Ιωάννα - 03115035

ΘΕΜΑ: Κωδικοποίηση Σημάτων Μουσικής βάσει Ψυχοακουστικού Μοντέλου.

Εισαγωγή

Σκοπός της συγκεκριμένης άσκησης είναι η συμπίεση του μουσικού σήματος music-dsp19.wav που περιέχεται στο συμπληρωματικό υλικό, κατανέμοντας σωστό αριθμό bits κβαντισμού ανά χρονικό τμήμα και κρίσιμη συχνοτική περιοχή, έτσι ώστε το λάθος κβαντισμού να είναι όσο το δυνατόν λιγότερο αντιληπτό καθώς επίσης και οι χρονο-συχνοτικές συνιστώσες του σήματος που ακούγονται περισσότερο να λαμβάνουν περισσότερο χώρο στην κωδικοποίηση από αυτές που επικαλύπτονται και χάνονται στη διαδικασία της ακοής. Η κατανομή των Bits κβαντισμού και η έμφαση στις αντιλήψιμες συχνότητες γίνεται αποτελεσματικότερα μελετώντας το σύστημα ακοής του ανθρώπου, πάνω στο οποίο βασίζεται το ψυχοακουστικό μοντέλο. Με βάση το μοντέλο αυτό, εισάγονται οι παρακάτω έννοιες:

Absolute Threshold of Hearing

Το *κατώφλι ακοής* χαρακτηρίζει το ελάχιστο ποσό της ενέργειας σε dB-Sound Pressure Level (dB SPL) που πρέπει να έχει ένας τόνος συχνότητας f έτσι ώστε να γίνει αντιληπτός από τον άνθρωπο σε περιβάλλον πλήρους ησυχίας. Η μη γραμμική συνάρτηση εκτίμησης του είναι η εξής:

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000 - 3.3)^2} + 10^{-3}(f/1000)^4 \text{ (dB SPL)} \quad (1)$$

Critical Bands

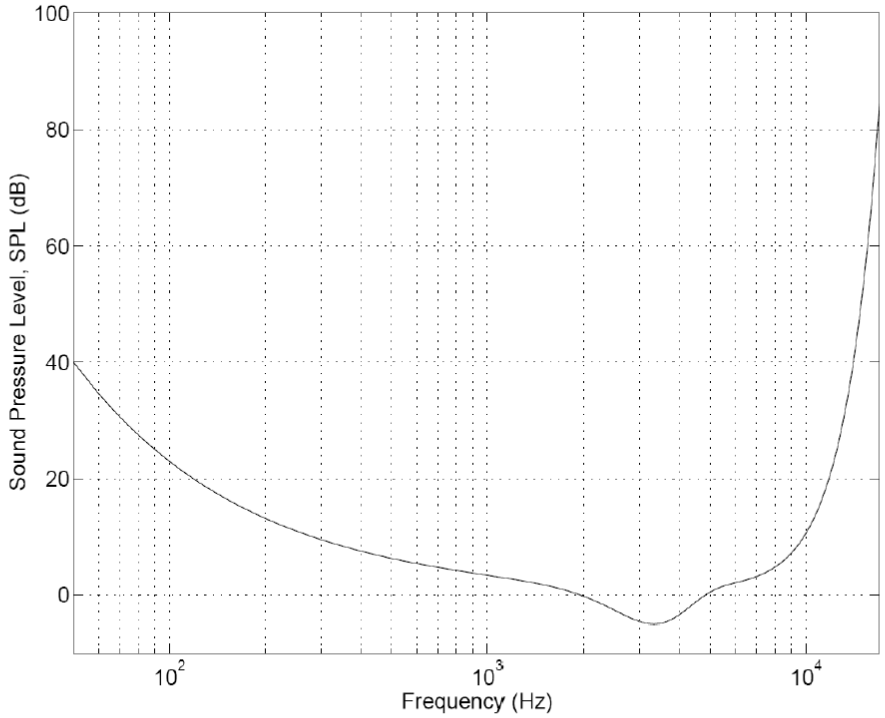
Οι *κρίσιμες συχνοτικές περιοχές* του ψυχοακουστικού μοντέλου έχουν σχεδιαστεί καθ' ομοίωση των περιοχών στις οποίες συντονίζονται οι νευροδέκτες του ακουστικού φλοιού. Οι 25 πρώτες κρίσιμες συχνοτικές περιοχές μοντελοποιούνται με την ψυχοακουστική κλίμακα συχνοτήτων bark, η οποία έχει πεδίο τιμών στο διάστημα [1,25]. Η μετατροπή των συχνοτήτων της κλίμακας Herz στην κλίμακα Bark βάσει του μη γραμμικού μοντέλου αντίληψης του ήχου γίνεται με τον τύπο:

$$b(f) = 13\tan^{-1}(0.00076f) + 3.5\tan^{-1}[(f/7500)^2] \text{ (Bark)} \quad (2)$$

Auditory Masking:

- Frequency Masking: Μια συνιστώσα σε μία συχνότητα f καλύπτει (masks) συνιστώσες σε γειτονικές συχνότητες με μικρότερη ένταση.
- Temporal Masking: Όταν δύο τόνοι παίζονται χρονικά κοντά, ο ένας καλύπτει τον άλλο.

Παρακάτω βλέπουμε ένα διάγραμμα με την τιμή του κατωφλιού ακοής (Absolute Threshold of Hearing) συναρτήσει των συχνοτήτων f του τόνου:



Σχήμα 1: Απόλυτη τιμή κατωφλίου ακοής: Absolute Threshold of Hearing.

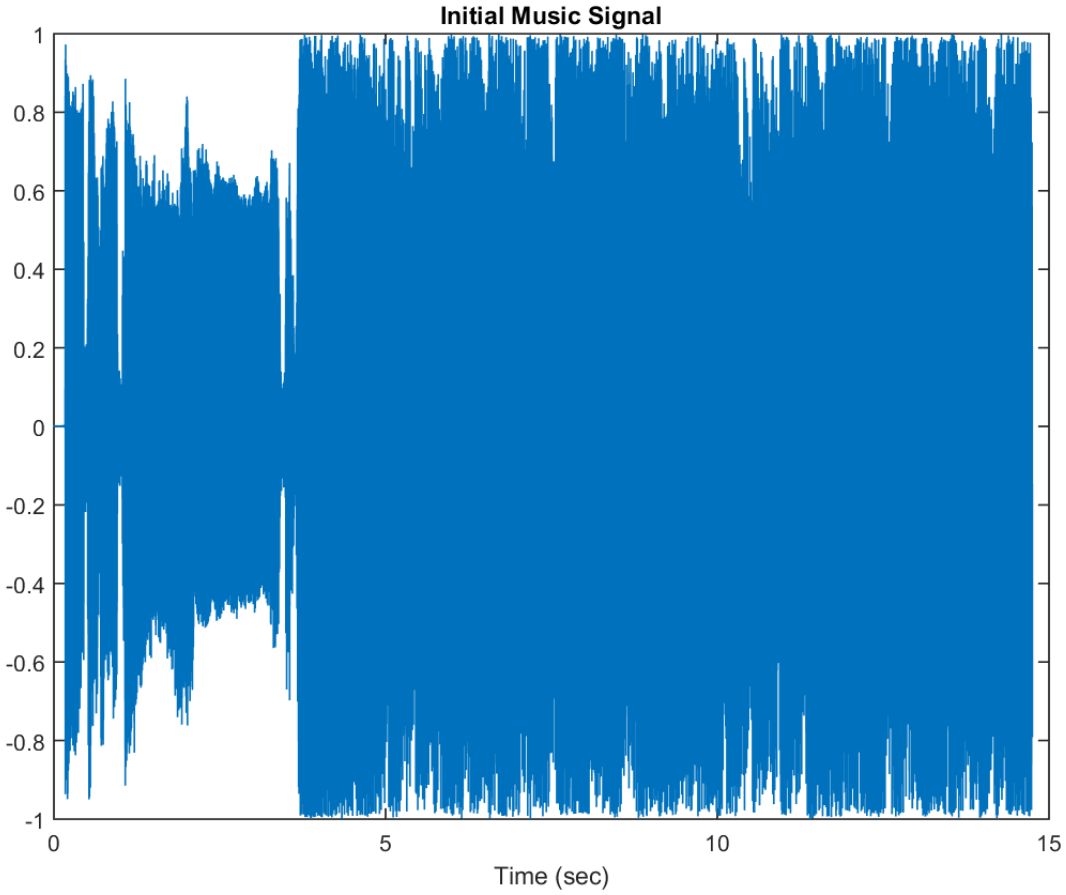
Παραθυροποίηση

Σύμφωνα με τα πρότυπα του MPEG-1 η ανάλυση του αρχικού σήματος γίνεται σε πλαίσια ανάλυσης x(n), με μήκος N = 512 δείγματα, τα οποία παραθυρώνονται με παράθυρο Hanning.

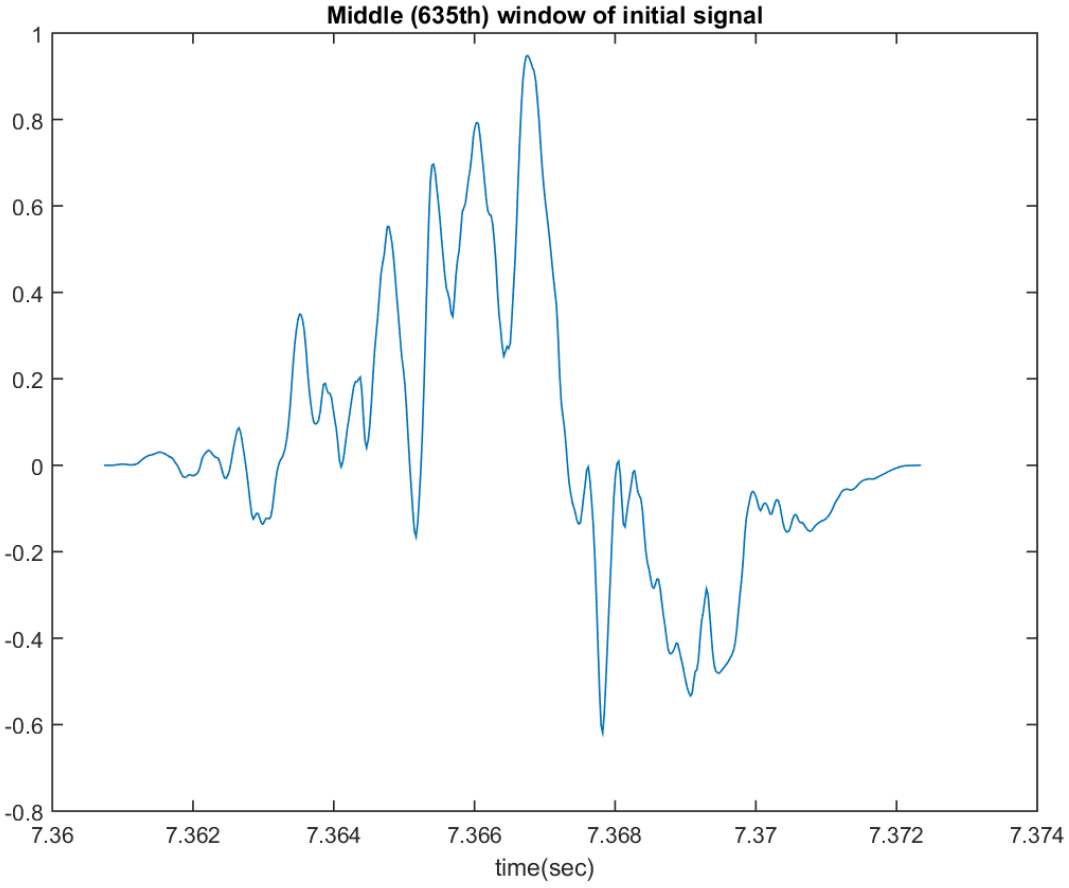
Μέρος 1ο – Ψυχοακουστικό Μοντέλο 1

Σκοπός του 1ου μέρους είναι η δημιουργία μίας συνάρτησης που υλοποιεί το ψυχοακουστικό μοντέλο 1, η οποία παίρνει σαν είσοδο το παραθυροποιημένο πλαίσιο ανάλυσης και επιστρέφει το συνολικό κατώφλι κάλυψης Tg.

Διαβάζουμε αρχικά το σήμα μουσικής με το MATLAB και κατόπιν το κανονικοποιούμε με την μέγιστη απόλυτη τιμή του, έτσι ώστε να έχει τιμές μεταξύ [-1,1]. Η γραφική παράσταση του κανονικοποιημένου σήματος στο πεδίο του χρόνου φαίνεται παρακάτω. Στη συνέχεια παραθυροποιούμε το σήμα και η ανάλυση γίνεται ξεχωριστά σε κάθε παράθυρο.



Η ανάλυση γίνεται σε κάθε παράθυρο, οι γραφικές παραστάσεις αφορούν όμως το μεσαίο παράθυρο (635ο παράθυρο – συνολικά έχουμε 1271) .



Στη συνέχεια υπολογίζουμε το φάσμα ισχύος P(k) του σήματος εκφρασμένο σε μονάδες SPL με τον τύπο:

$$P(k) = PN + 10\log_{10}\left|\sum_{n=0}^{N-1} w(n)x(n)e^{-j\frac{2\pi kn}{N}}\right|^2 \text{, όπου PN = 90.302 dB.}$$

Αφού υπολογιστεί το φάσμα ισχύος P(k), στη συνέχεια θέλουμε να εντοπίσουμε ανά critical band τοπικά μέγιστα (μάσκες) τα οποία είναι μεγαλύτερα από τις γειτονικές τους συχνότητες τουλάχιστον κατά 7 dB (Frequency Masking). Το εύρος της γειτονιάς υπολογισμού των μασκών διαφέρει ανάλογα με την διακριτή συχνότητα k. Στις υψηλές συχνότητες οι μάσκες καλύπτουν ευρύτερες γειτονιές. Ορίζουμε έτσι την συνάρτηση St(k), η οποία επιστρέφει λογικές τιμές {0,1} που προσδιορίζουν αν στην θέση k υπάρχει τονική μάσκα (τιμή 1). Η συνάρτηση καθορίζει την ύπαρξη μάσκας προσδιορίζοντας τοπικά μέγιστα στις διαφορετικές συχνοτικές περιοχές όπως ορίζονται παρακάτω:

$$St(k) = \begin{cases} 0, \alpha\nu\ k\ \notin [3,250] \\ P(k) > P(k \pm 1)\ \wedge P(k) > P(k \pm \Delta\kappa) + 7dB, \alpha\nu\ k\ \in [3,250] \end{cases}$$

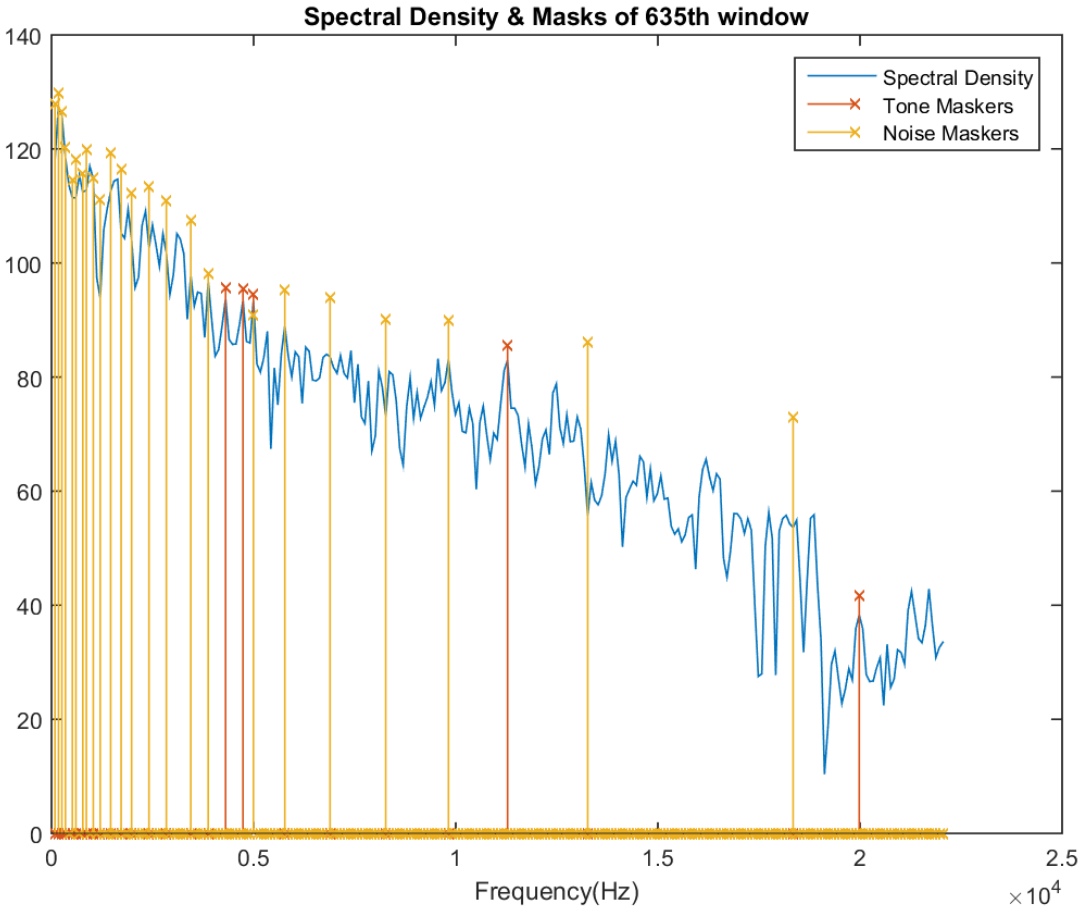
,όπου

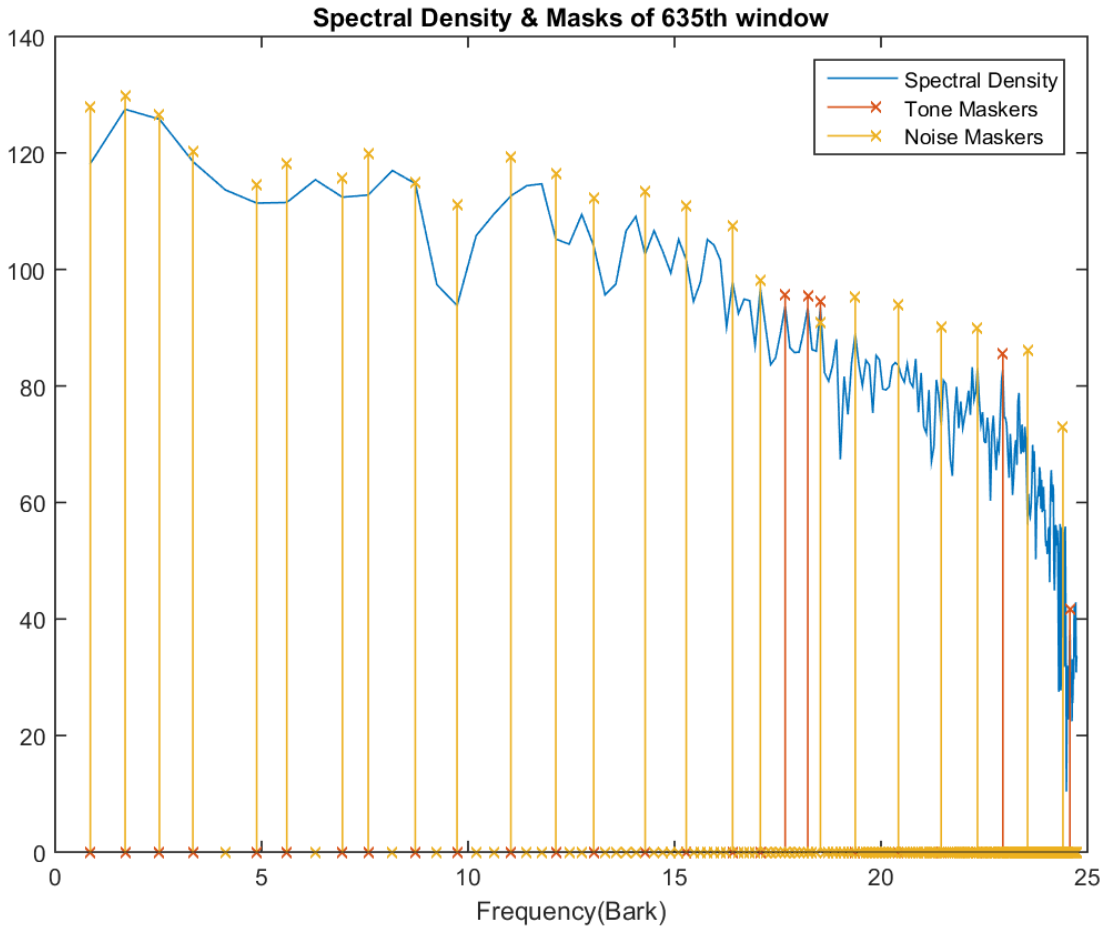
$$\Delta\kappa = \begin{cases} 2\ ,\ 2 < \kappa < 63 \\ [2,3],\ 63 \leq \kappa < 127 \\ [2,6]\ ,\ 127 \leq \kappa\ \leq 250 \end{cases}$$

Μόλις βρεθούν οι θέσεις των τονικών μασκών υπολογίζουμε την ισχύ τους $P_{TM}(k)$ ως εξής:

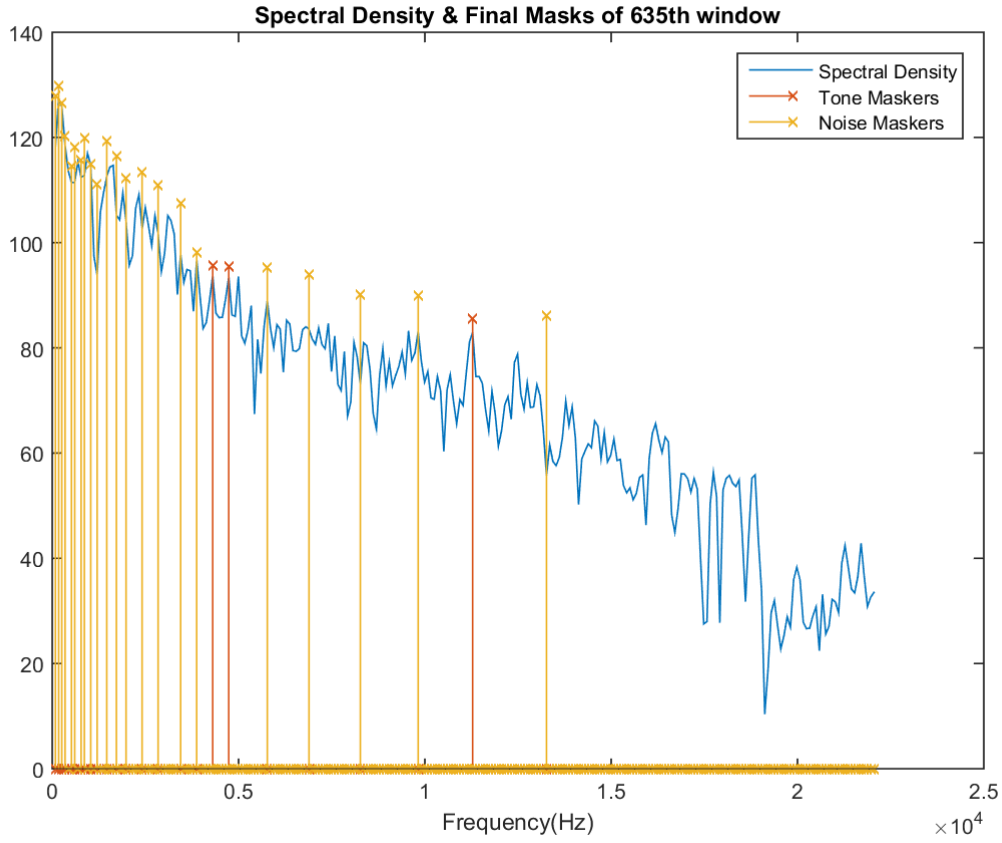
$$P_{TM}(k) = \begin{cases} 0\ , \alpha\nu\ St(k) = 0 \\ 10log_{10}\left(10^{0.1P(k-1)} + 10^{0.1P(k)} + 10^{0.1P(k+1)}\right),\ \ \alpha\nu\ St(k) = 1 \end{cases}$$

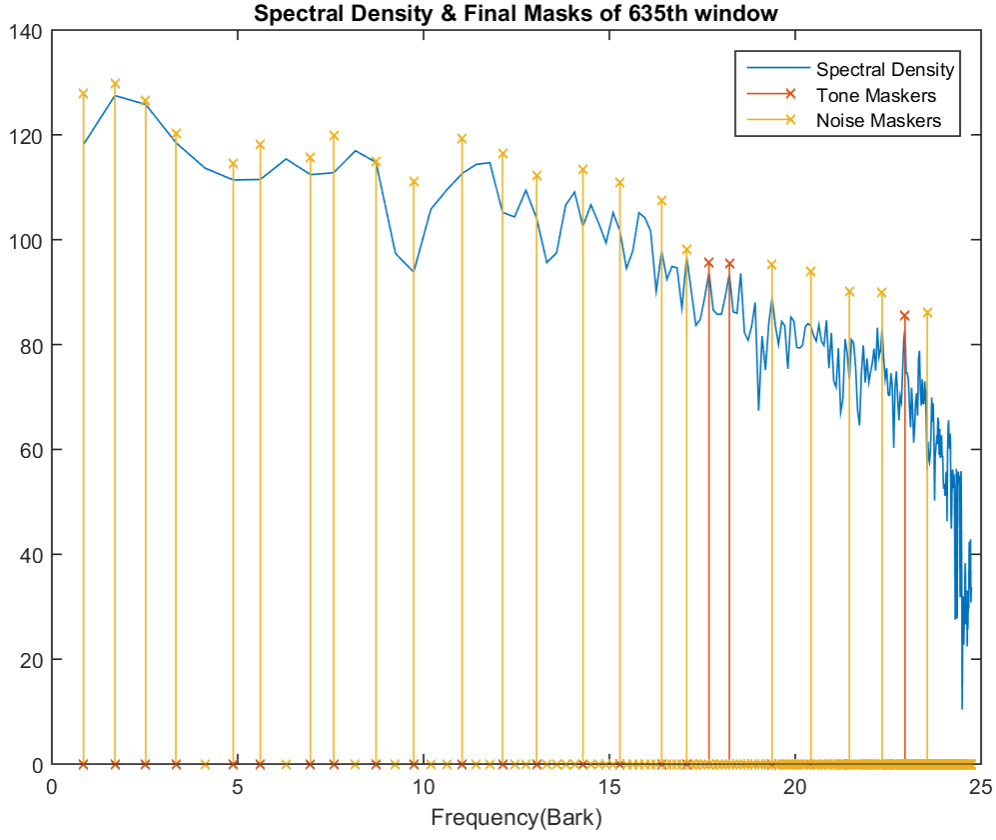
Στη συνέχεια για την εύρεση των μασκών που οφείλονται στην ύπαρξη θορύβου χρησιμοποιούμε την έτοιμη συνάρτηση που περιέχεται στο συμπληρωματικό υλικό findNoiseMaskers(P, P_{TM}, b) , όπου b είναι οι συχνότητες εκφρασμένες σε κλίμακα bark όπως ορίζεται από την εξίσωση (2). Παριστάνοντας γραφικά τα τρία παραπάνω μεγέθη έχουμε τις εξής παραστάσεις (σε κλίμακα Herz και Bark αντίστοιχα):


Spectral Density & Masks of 635th window
Spectral Density
Tone Maskers
Noise Maskers
140
120
100
80
60
40
20
0
0
0.5
1
1.5
2
2.5
Frequency(Hz)
×10^4


Spectral Density & Masks of 635th window
Spectral Density
Tone Maskers
Noise Maskers
140
120
100
80
60
40
20
0
0
5
10
15
20
25
Frequency(Bark)

Στη συνέχεια χρησιμοποιώντας τη δοσμένη συνάρτηση checkMaskers(P_{TM}, P_{NM}, T_q, b) και με βάση δύο διαφορετικά κριτήρια μειώνουμε τον αριθμό των μασκών. Έτσι οι τελικές μάσκες τόνων και θορύβου είναι οι εξής (σε κλίμακες Herz και Bark):





Έπειτα αφού υπολογίσουμε την έκταση κάλυψης SF(i,j) μίας μάσκας (τονικής και θορύβου ξεχωριστά) που βρίσκεται στο σημείο j για όλα τα σημεία i, υπολογίζουμε τα ξεχωριστά κατώφλια κάλυψης T_{TM} και T_{NM}. Το κάθε κατώφλι αντιπροσωπεύει το ποσοστό κάλυψης στο σημείο i από την μάσκα τόνου ή θορύβου αντίστοιχα στο σημείο j. Τα δύο κατώφλια υπολογίζονται ως εξής:

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

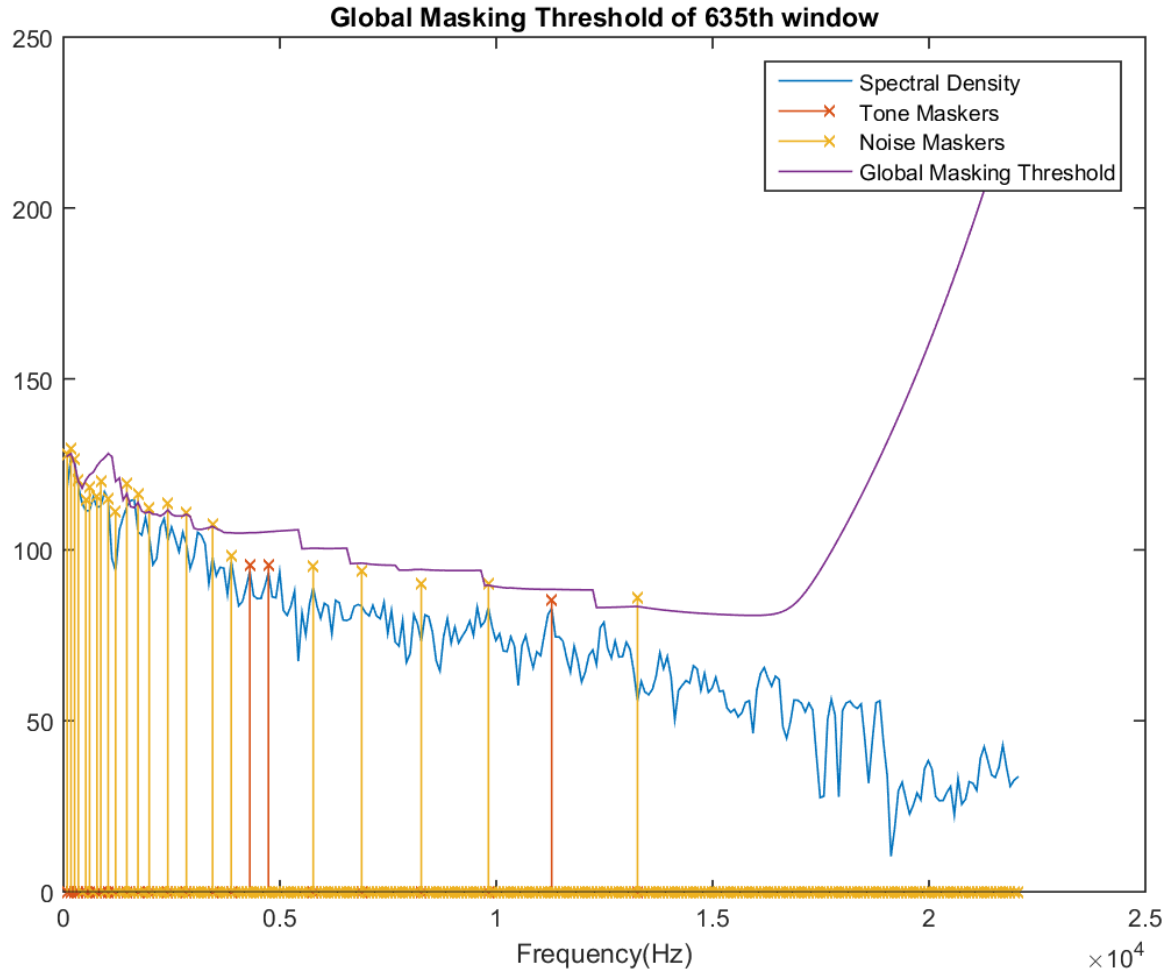
$$T_{NM}(i,j) = P_{NM}(j) - 0.175b(j) + SF(i,j) - 2.025 \text{ (dB SPL)}$$

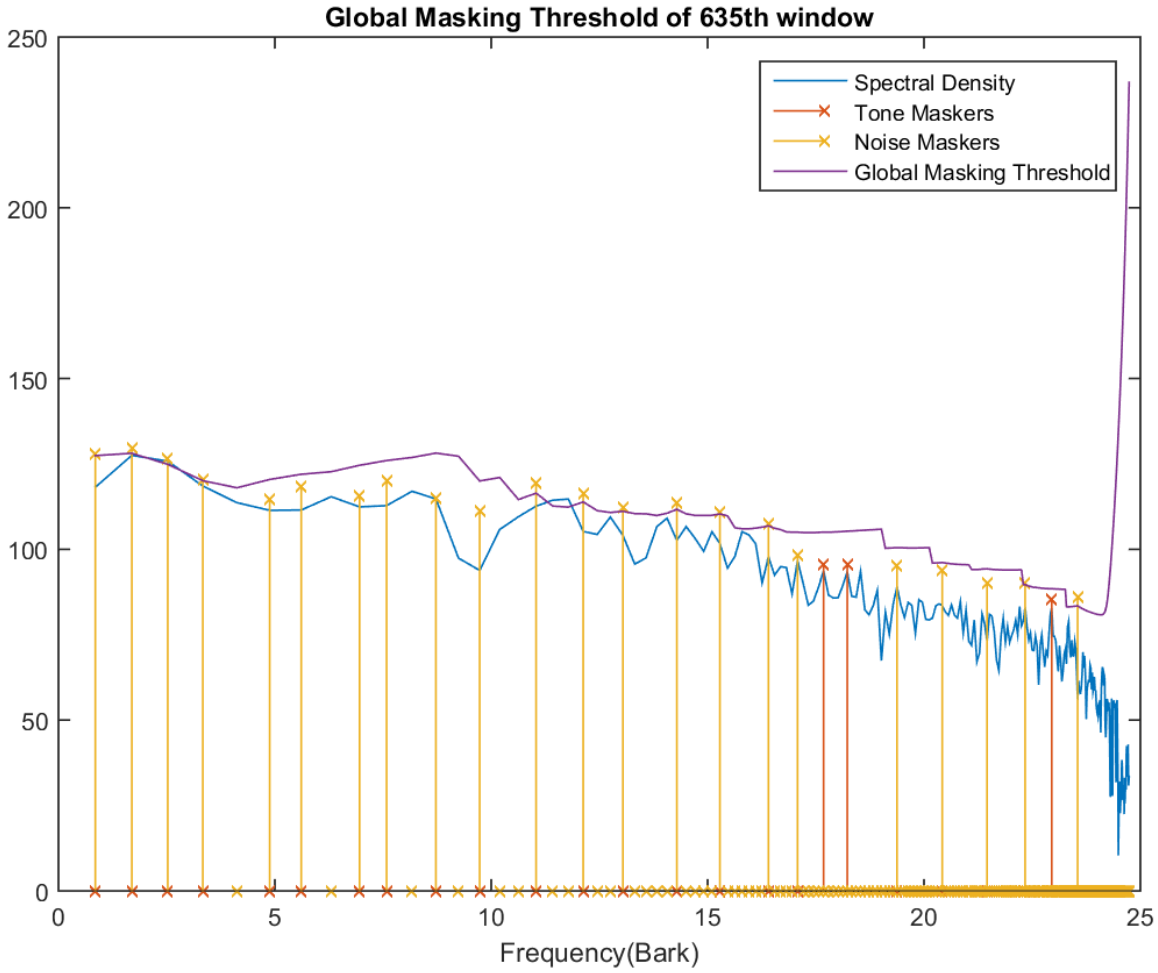
Στη συνέχεια υπολογίζουμε το συνολικό κατώφλι κάλυψης σε κάθε διακριτή συχνότητα, συνδυάζοντας τα ξεχωριστά κατώφλια κάλυψης που υπολογίστηκαν παραπάνω (το Absolute Threshold of Hearing (Tq), τα κατώφλια κάλυψης των τόνων T_{TM} και του θορύβου T_{NM}). Το συνολικό κατώφλι κάλυψης Tg(i) σε κάθε διακριτή συχνότητα i υπολογίζεται ως εξής:

$$Tg(i) = 10\log_{10}(10^{0.1Tq(i)} + \sum_{l=1}^{L} 10^{0.1T_{TM}(i,l)} + \sum_{m=1}^{M} 10^{0.1T_{NM}(i,m)}) \text{ Db SPL}$$

,όπου L,M το μήκος ο τελικός αριθμός των μασκών τόνων και θορύβου αντίστοιχα.

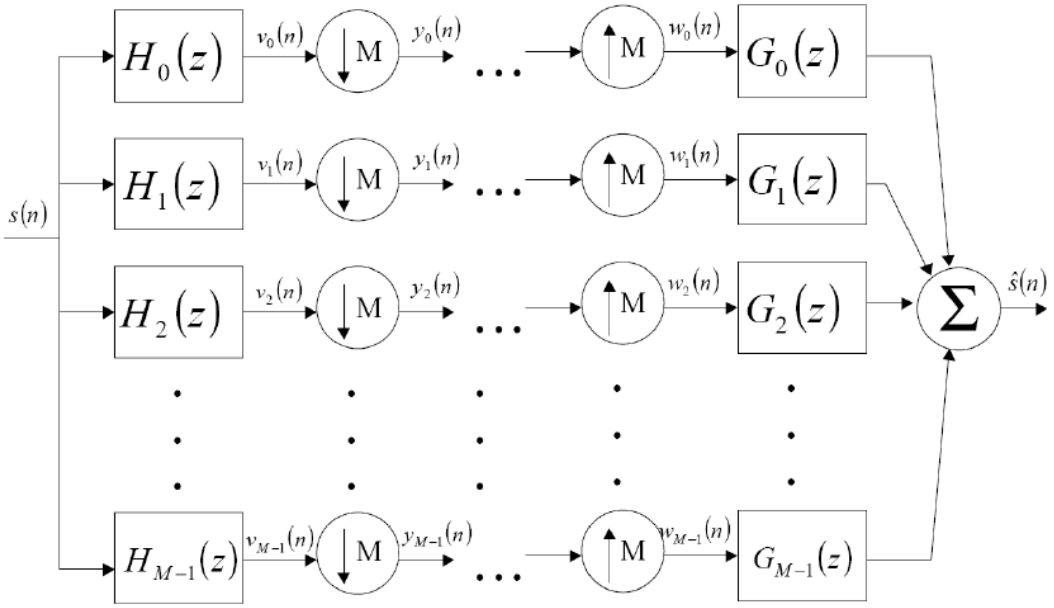
Έτσι τελικά βρίσκουμε την ελάχιστη ένταση που πρέπει να έχει ένας τόνος του συγκεκριμένου σήματος μουσικής ώστε να γίνει αντιληπτός από τον άνθρωπο. Το αποτέλεσμα φαίνεται στην παρακάτω γραφική παράσταση:





Μέρος 2ο – Χρονο-Συχνοτική Ανάλυση με Συστοιχία Ζωνοπερατών Φίλτρων

Στο μέρος αυτό σκοπός είναι η κβαντοποίηση του σήματος μουσικής, χρησιμοποιώντας για κάθε παράθυρο του αρχικού σήματος το συνολικό κατώφλι κάλυψης που βρέθηκε στο 1ο μέρος, καθώς και συστοιχίες ζωνοπερατών φίλτρων, οι οποίες χρησιμοποιούνται για την χρονο-συχνοτική ανάλυση του σήματος. Συγκεκριμένα διαιρούν το φάσμα σε υποζώνες συχνοτήτων και με αυτόν τον τρόπο παρέχονται πληροφορίες σχετικά με την συχνοτική κατανομή του σήματος, οι οποίες βοηθούν στην ταυτοποίηση των αντιληπτικά περιττών σημείων του σήματος. Στόχος λοιπόν είναι η δημιουργία μίας συνάρτησης που υλοποιεί την διαδικασία του παρακάτω σχήματος , η οποία παίρνει σαν είσοδο κάθε πλαίσιο ανάλυσης x(n), τη συστοιχία φίλτρων και το συνολικό κατώφλι κάλυψης (1ο μέρος) και επιστρέφει το ανακατασκευασμένο σήμα $\hat{x}(n)$ καθώς και τον αριθμό των bits ανά δείγμα που χρησιμοποιήθηκαν για τη δημιουργία του.



Σχήμα 3: Uniform M-Band Maximally Decimated Analysis-Synthesis Filterbank.

Αρχικά κατασκευάζουμε την συστοιχία των ζωνοπερατών φίλτρων ανάλυσης και σύνθεσης, $h_k(n)$ και $g_k(n)$. Στο στάδιο κωδικοποίησης και αποκωδικοποίησης του συστήματος συμπίεσης που υλοποιούμε, χρησιμοποιούνται M = 32 φίλτρα ανάλυσης και σύνθεσης αντίστοιχα, των οποίων οι κρουστικές αποκρίσεις υπολογίζονται ως εξής:

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

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

,όπου L=2M=64 και $\sin\left[\left(n+\frac{1}{2}\right)\frac{\pi}{2M}\right]$ (για 0≤n≤L-1 και 0≤k≤M-1) ένα βαθυπερατό φίλτρο γνωστό ως ημιτονικό παράθυρο.

Στη συνέχεια κάνουμε συνέλιξη κάθε παραθύρου του αρχικού σήματος με κάθε ένα από τα 32 φίλτρα ανάλυσης $h_k(n)$ και υποδειγματοληπτούμε το φιλτραρισμένο σήμα κατά ακέραιο παράγοντα M ώστε να διαιρεθεί το αρχικό σήμα στις χρονικές του συνιστώσες. Ο αποδεκατισμός που επιδέχεται γίνεται σε μέγιστο βαθμό ώστε να μην υπάρχουν φαινόμενα επικάλυψης (aliasing). Ωστόσο, τα μη ιδανικά ζωνοπερατά φίλτρα εισάγουν επικαλύψεις μεταξύ των συνιστωσών, τις οποίες θεωρούμε αμελητέες.

Έπειτα προχωράμε στην υλοποίηση του κβαντιστή, δηλαδή στην αντιστοίχιση των δειγμάτων του σήματος σε αριθμημένα επίπεδα κβάντισης. Για την υλοποίηση της άσκησης θα χρησιμοποιήσουμε έναν ομοιόμορφο προσαρμοζόμενο κβαντιστή 2^{B_κ} επιπέδων, όπου B_κ ο αριθμός των bits ανά δείγμα για κάθε φιλτραρισμένη και υποδειγματοληπτημένη ακολουθία $y_k(n)$ του τρέχοντος παραθύρου ανάλυσης του σήματος. Το βήμα Δ του κβαντιστή προσαρμόζεται σε κάθε πλαίσιο ανάλυσης και για κάθε τέτοια ακολουθία βάσει του B_κ έτσι ώστε : $\Delta = \frac{|x_{max}-x_{min}|}{B_\kappa}$, όπου x_{max},x_{min} οι μέγιστη και η ελάχιστη τιμή του τρέχοντος παραθύρου του σήματος. Τα επίπεδα του κβαντιστή πρέπει να είναι αρκετά έτσι ώστε το σφάλμα κβαντισμού να μη γίνεται αντιληπτό από τον άνθρωπο μετά την συμπίεση του σήματος μουσικής. Το μέγιστο ανεκτό σφάλμα συνδέεται με το συνολικό κατώφλι κάλυψης Tg(i) του ψυχοακουστικού μοντέλου όπως προέκυψε από το 1ο μέρος και η σχέση υπολογισμού του είναι η εξής:

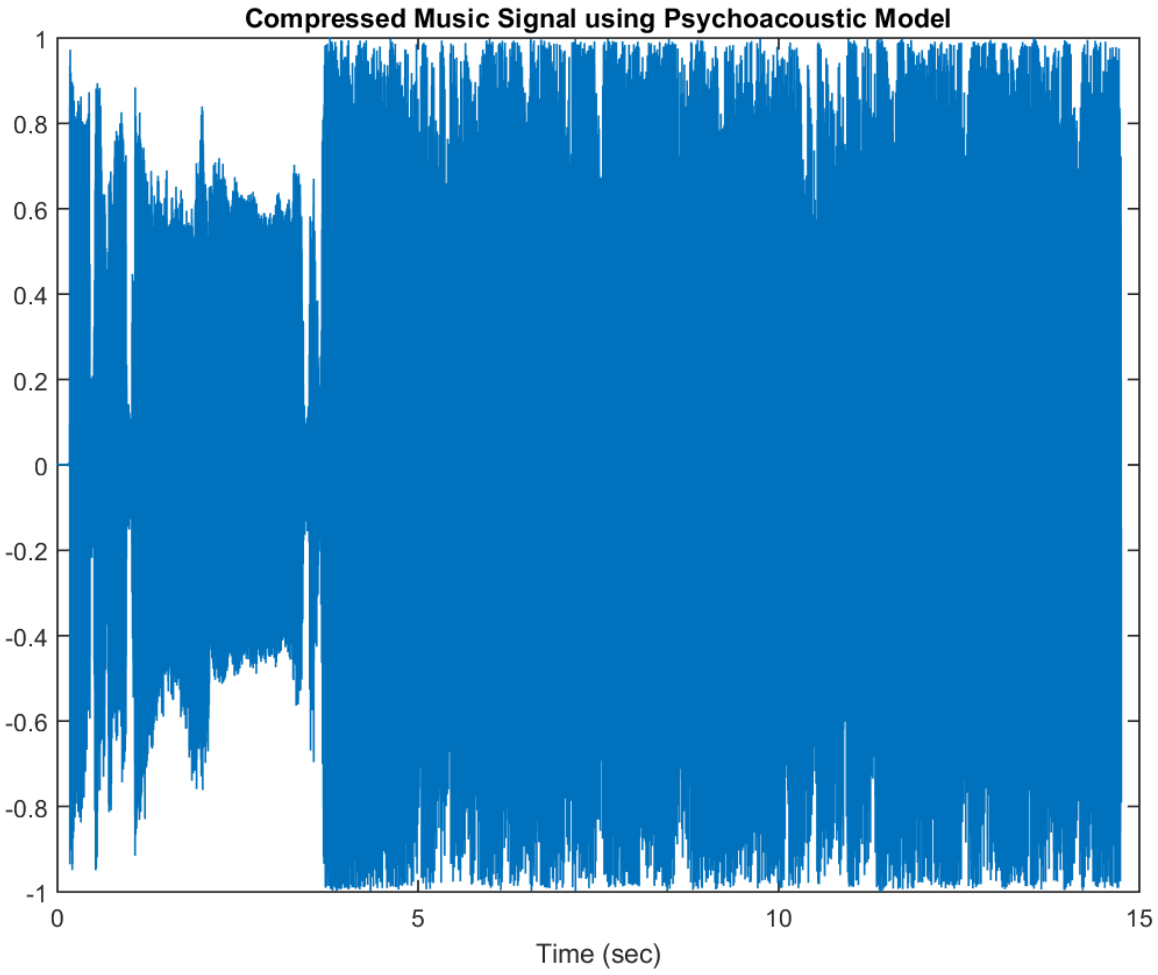
$$B_k = \log_2\left(\frac{R}{min_{8(k-1)+1\leq 8k}(Tg(i))}\right) - 1$$

,όπου R ο αριθμός των βαθμίδων έντασης του αρχικού σήματος (εδώ R = 2^{16}) . Στον προσαρμοζόμενο κβαντιστή για μικρότερο λάθος κβαντισμού, η θέση του πρώτου επιπέδου προσαρμόζεται με βάση την ελάχιστη τιμή του σήματος x_{min} σε κάθε πλαίσιο ανάλυσης. Στη συνέχεια οι κβαντισμένες ακολουθίες $\widehat{y_k}(n)$ στέλνονται στον αποκωδικοποιητή, όπου παρεμβάλλονται M μηδενικά και υπερδειγματοληπτούνται σύμφωνα με την σχέση:

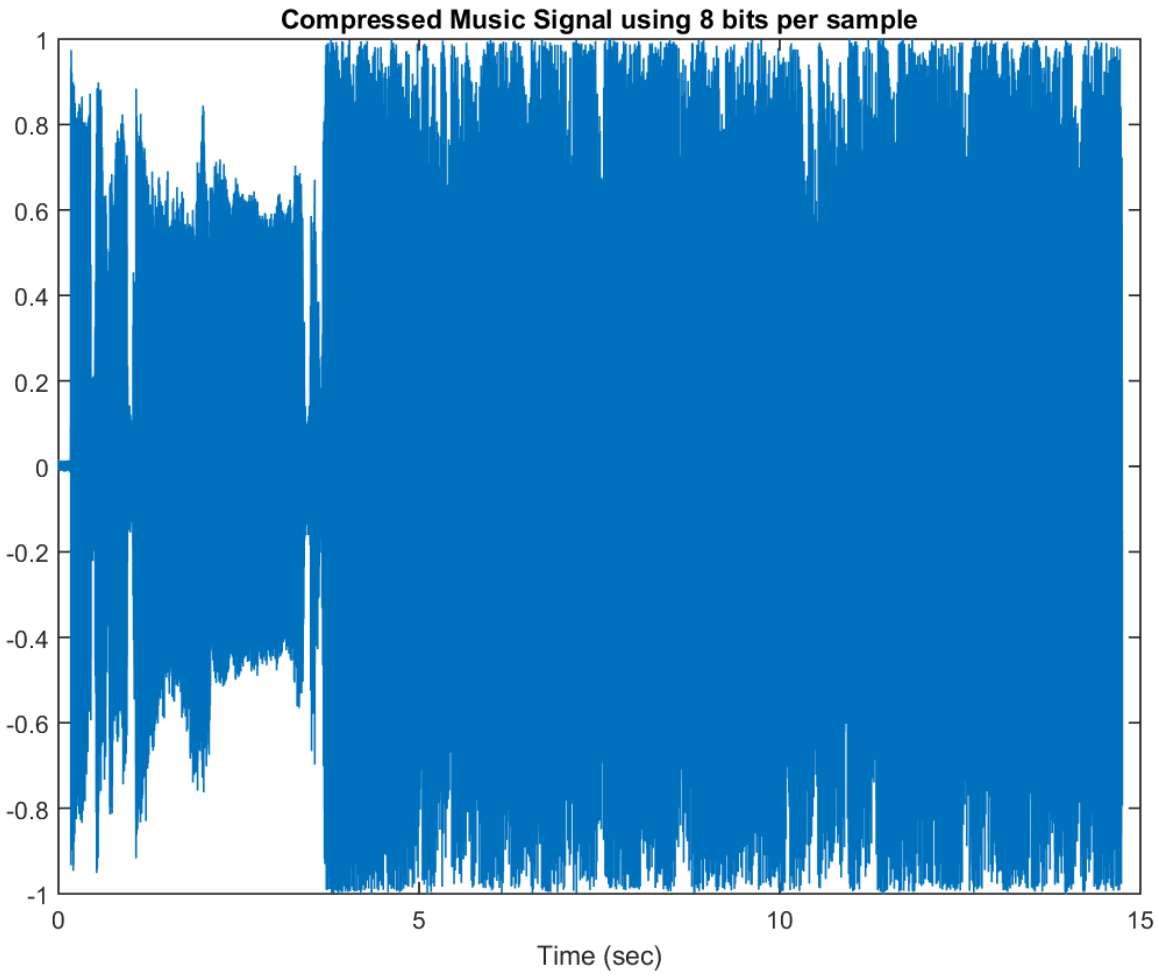
$$W_k(n) = \begin{cases} \widehat{y_k}\left(\frac{n}{M}\right), n = 0, M, 2M, 3M \ldots \\ 0, \ \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$$

Έπειτα οι ακολουθίες $w_k(n)$ συνελίσοτνται με τα φίλτρα σύνθεσης $g_k(n)$,τα οποία λόγω της αντιστροφής τους ως προς τον χρόνο ικανοποιούν τον περιορισμό ως προς τη γραμμική φάση. Η τελική ανακατασκευή του σήματος μουσικής γίνεται με την εφαρμογή της τεχνικής Overlap-Add, χωρίς επικάλυψη μεταξύ των διαδοχικών πλαισίων ανάλυσης.

Έτσι το συμπιεσμένο σήμα με χρήση του ψυχοακουστικού μοντέλου είναι το εξής:

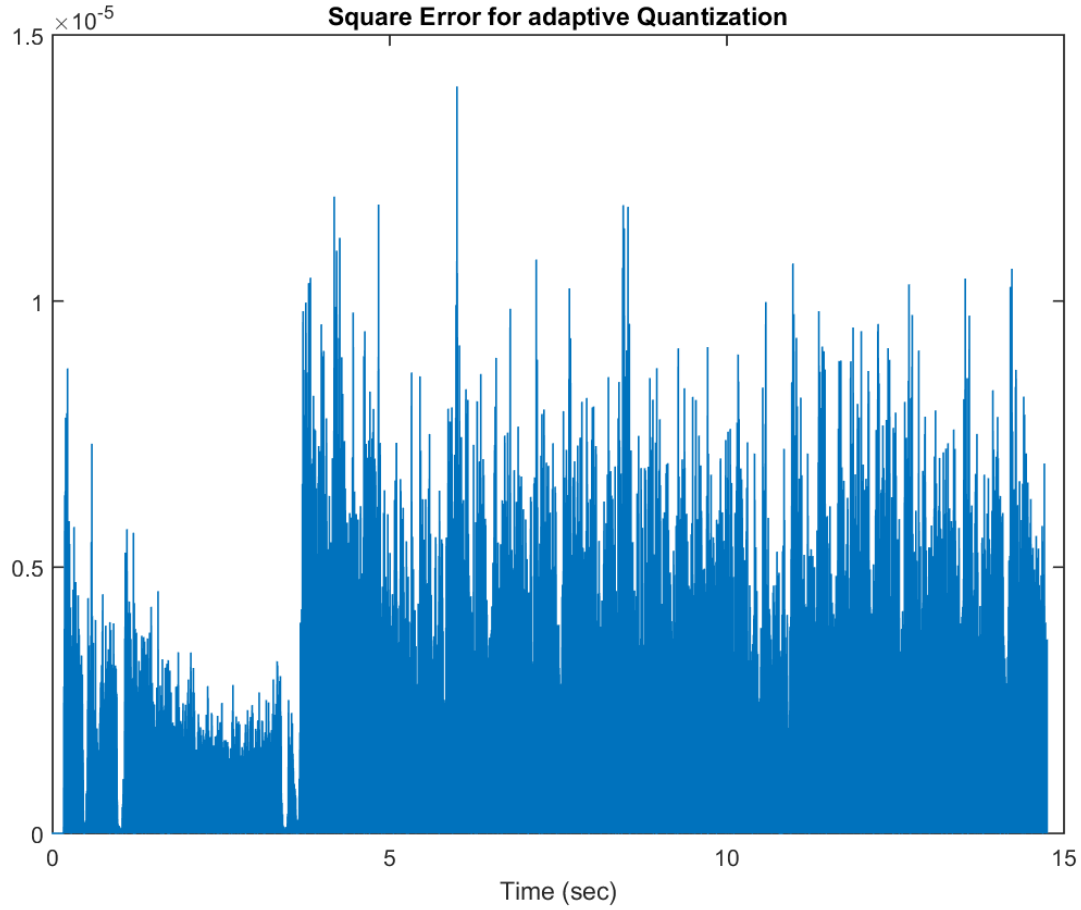


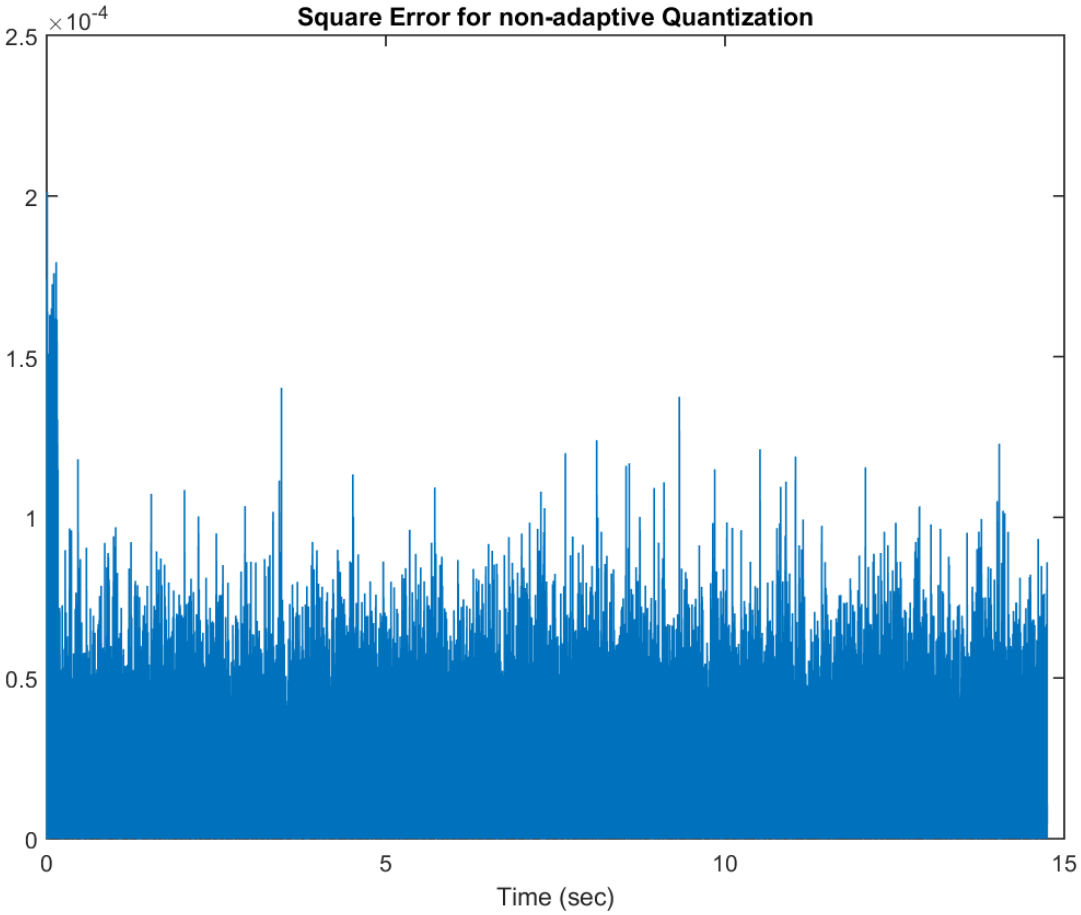
Στη συνέχεια πειραματιστήκαμε την υλοποίηση ενός απλού, μη-προσαρμοζόμενου κβαντιστή με σταθερό αριθμό bit ανά δείγμα, $B_κ$ = 8 bits και θεωρώντας $[x_{min}\ x_{max}]$ = [-1 1] σταθερά για κάθε παράθυρο ανάλυσης του σήματος. Το συμπιεσμένο σήμα είναι το εξής:



Αποτελέσματα συγκρίσεων

Για την σύγκριση των δύο παραπάνω μεθόδων χρησιμοποιούμε δύο μεγέθη: i) τα ποσοστά συμπίεσης (Compression Ratio) των τελικών σημάτων ως προς το αρχικό και ii) το μέσο τετραγωνικό σφάλμα (Mean Square Error – MSE), το οποίο ορίζεται ως η μέση τιμή του τετραγώνου της διαφοράς του αρχικού από το ανακατασκευασμένου σήμα μουσικής. Έτσι για τον προσαρμοζόμενο κβαντιστή βρήκαμε ποσοστό συμπίεσης ίσο με 62.56%, ενώ για τον απλό βρήκαμε 56.28%. Όσον αφορά το μέσο τετραγωνικό σφάλμα προέκυψε ότι στην περίπτωση του προσαρμοζόμενου κβαντιστή έχουμε MSE = $7.8235*10^{-7}$, ενώ στην άλλη περίπτωση MSE = $6.5602*10^{-6}$. Είναι προφανές λοιπόν ότι ο προσαρμοζόμενος κβαντιστής πετυχαίνει μεγαλύτερη συμπίεση του αρχικού σήματος, παρουσιάζοντας παράλληλα μικρότερη απόκλιση από το αρχικό σήμα. Το λάθος που προκύπτει για τις δύο περιπτώσεις ξεχωριστά σε κάθε χρονική στιγμή φαίνονται παρακάτω:


Square Error for adaptive Quantization
×10^-5
1.5
1
0.5
0
0
5
10
15
Time (sec)


Square Error for non-adaptive Quantization
×10^-4
2.5
2
1.5
1
0.5
0
0
5
10
15
Time (sec)