Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Ψηφιακή Επεξεργασία Σήματος

2η εργαστηριακή άσκηση

«Κωδικοποίηση σημάτων Μουσικής βάσει ψυχοακουστικού μοντέλου (Perceptual Audio Coding) »

Βασιλοπούλου Φωτεινή ΑΜ: 03114854

Ακαδημαϊκό έτος 2018-2019

Εισαγωγή

Στόχος της εργασίας είναι η δημιουργία δύο συναρτήσεων. Η πρώτη συνάρτηση *PsychoacousticModel( )* υλοποιεί το Ψυχοακουστικό Μοντέλο Ι παράγοντας το συνολικό κατώφλι κάλυψης T_g. Στη συνέχεια, σκοπός είναι η χρονο-συχνοτική ανάλυση του σήματος μουσικής χρησιμοποιώντας συστοιχία ζωνοπερατών φίλτρων. Το δεύτερο αυτό τμήμα της εργασίας πραγματοποιείται με τη συνάρτηση *AnalysisSynthesisFilterbank( ),* η οποία επιστρέφει το ανακατασκευασμένο σήμα μουσικής σύμφωνα με τον τρόπο κβάντισης που έχουμε επιλέξει. Η επεξεργασία του αρχικού σήματος μουσικής, η κλήση των δύο συναρτήσεων που δημιουργήθηκαν και η εξαγωγή ορισμένων συγκριτικών μεγεθών πραγματοποιείται συνολικά στο αρχείο *code_dsp2019( ).*

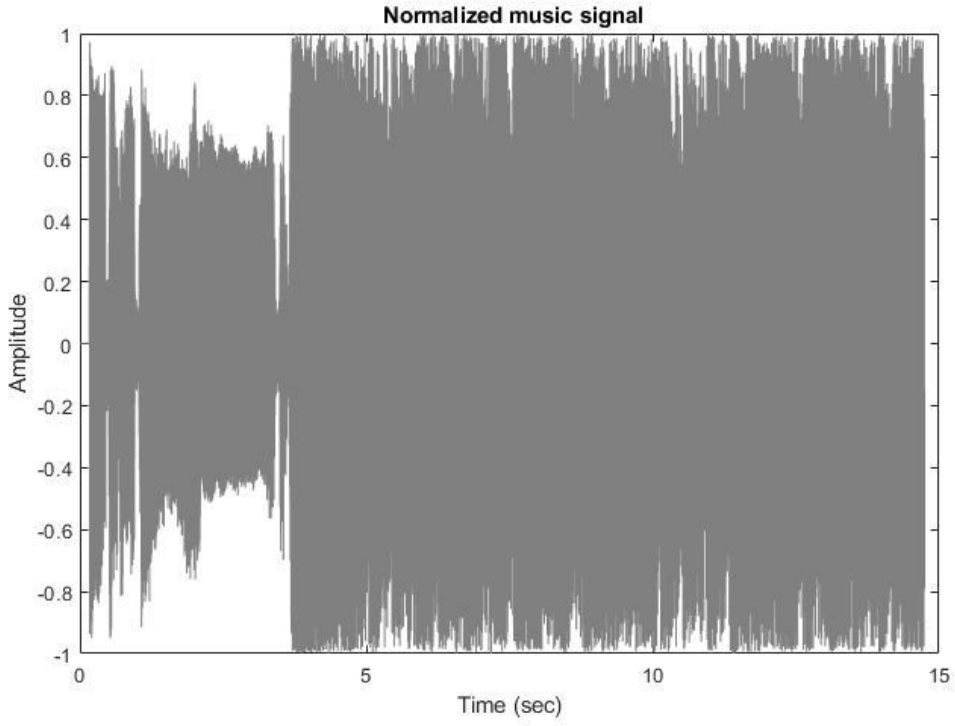
Μέρος 1. Ψυχοακουστικό Μοντέλο 1

Βήμα 1.0 : Κανονικοποίηση του σήματος

Το αρχείο μουσικής που περιλαμβάνεται στα υλικά της άσκησης αποτελείται από δύο πίνακες-στήλες που περιλαμβάνουν το ίδιο σήμα από δύο διαφορετικά κανάλια. Το σήμα μουσικής που χρησιμοποιείται στη συνέχεια για την επεξεργασία λαμβάνεται ως ο μέσος όρος των δύο αυτών σημάτων, δηλαδή

$$\mathrm{x[n]} = \frac{1}{2}(x_L[n] + x_R[n])$$

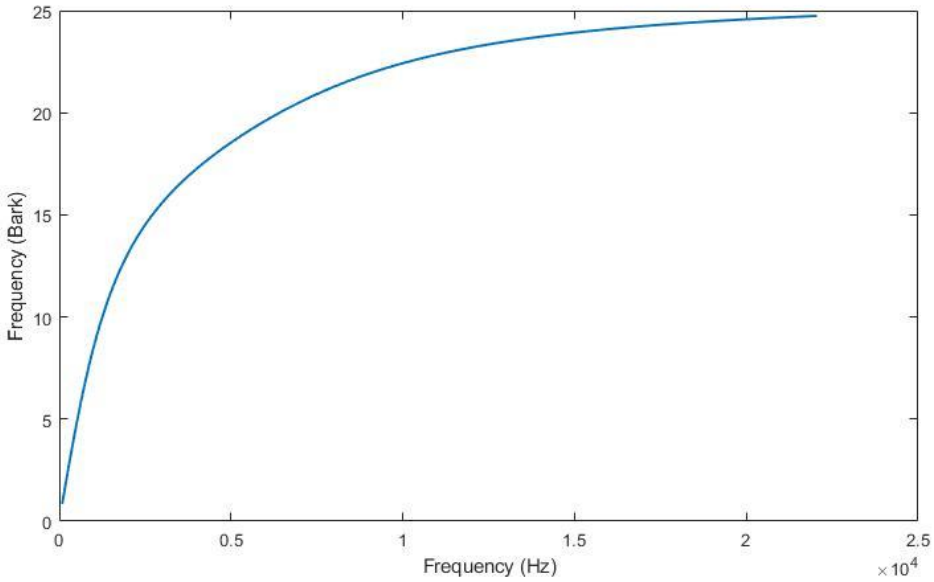
Αρχικά κανονικοποιούμε το σήμα μουσικής διαιρώντας το με την απόλυτη τιμή της μέγιστης τιμής του, ώστε τελικά το πλάτος του να είναι στο διάστημα [-1,1]. Το σήμα μουσικής φαίνεται στην *Εικόνα 1* που ακολουθεί. Επίσης πραγματοποιείται παραθυροποίηση του σήματος μουσικής για την μετέπειτα επεξεργασία του. Κάθε παράθυρο αποτελείται σύμφωνα με τις οδηγίες από *N=512* δείγματα.



Εικόνα 1: Κανονικοποιημένο σήμα μουσικής σε συνάρτηση με το χρόνο.

Επίσης μετατρέπουμε τις συχνότητες της κλίμακας *Hertz* στην κλίμακα *Bark* σύμφωνα με τη σχέση:

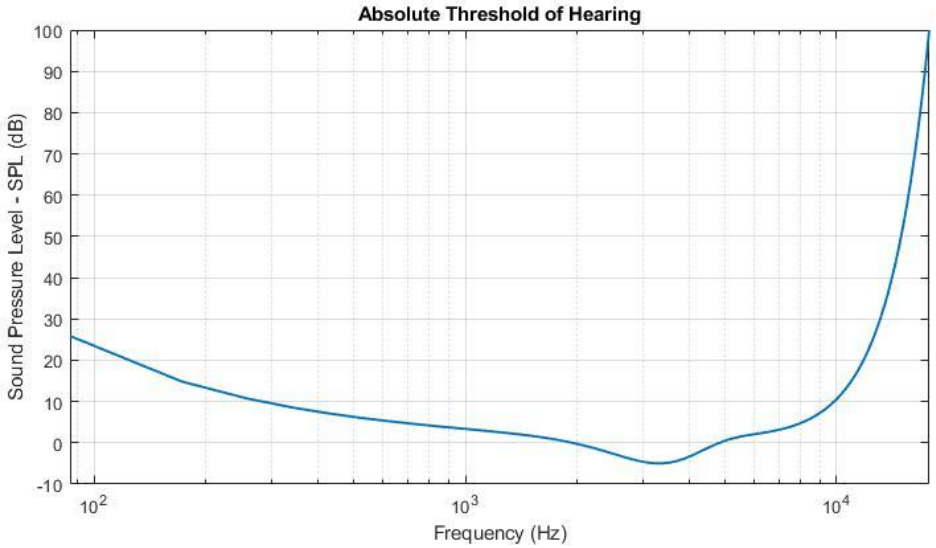
$$b(f) = 13\tan^{-1}(0.00076f) + 3.5\tan^{-1}\left[\left(\frac{f}{7500}\right)^2\right] \quad (Bark)$$



Εικόνα 2: Μετατροπή συχνοτήτων από την κλίμακα Hz στην κλίμακα Bark.

Επίσης δημιουργούμε το Κατώφλι Ακοής *(Absolute Threshold of Hearing)* το οποίο αντιπροσωπεύει το ποσό της ενέργειας σε *Sound Pressure Level (SPL dB)* που πρέπει να έχει ένας τόνος συχνότητας *f* ώστε να γίνει αντιληπτός από τον άνθρωπο σε περιβάλλον πλήρους ησυχίας. Η τιμή αυτή, για κάθε τιμή συχνότητας, μπορεί να ερμηνευθεί ως το μέγιστο όριο ενέργειας για την κωδικοποίηση παραμορφώσεων που εισάγονται στο πεδίο της συχνότητας. Το Κατώφλι Ακοής ορίζεται από τη σχέση:

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



Εικόνα 3: Απόλυτη τιμή κατωφλίου ακοής (Absolute Threshold of Hearing)

Βήμα 1.1 : Φασματική Ανάλυση

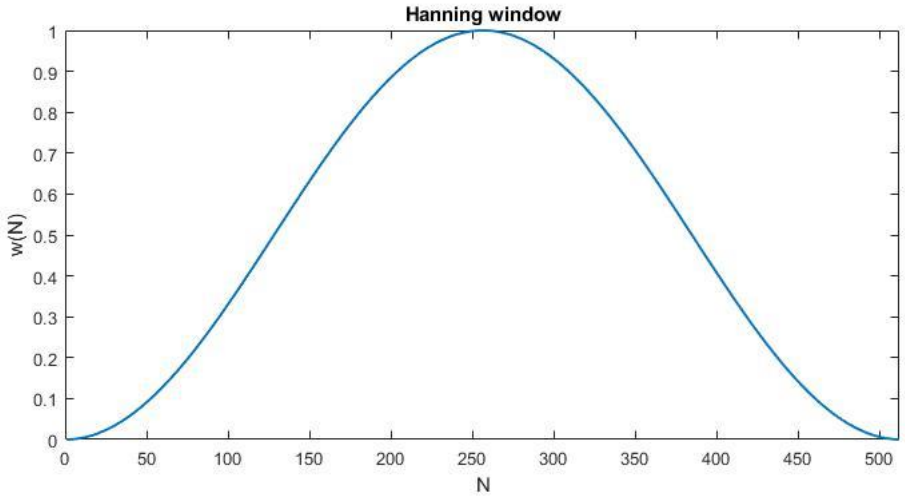
Επιλέγουμε να παρουσιάσουμε τα αποτελέσματα της επεξεργασίας για το παράθυρο 984 του σήματος μουσικής.

Για κάθε πλαίσιο του αρχικού σήματος μουσικής υπολογίζουμε το N-σημείων φάσμα ισχύος $P(k)$ σε μονάδες *SPL (Sound Pressure Level)* χρησιμοποιώντας παράθυρα *Hanning*.

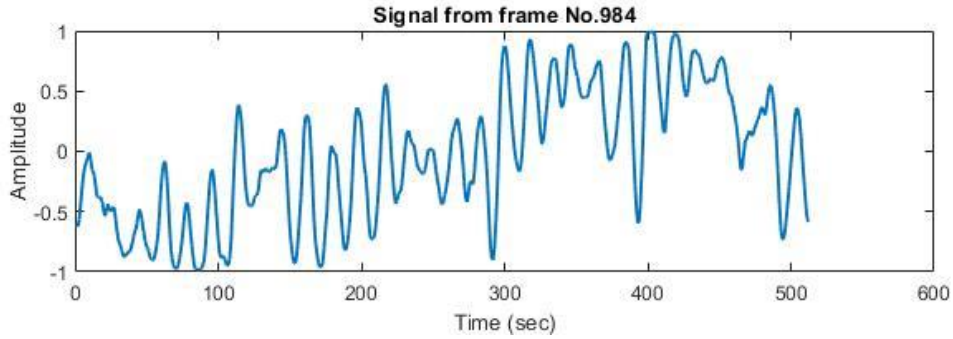
$$P(k) = 90.302 + 10\log_{10}\left|\sum_{n=0}^{N-1} w(n)x(n)e^{-j\frac{2\pi kn}{N}}\right|^2 \ (dB)\ , \qquad 0 \le k \le \frac{N}{2}$$

Τα παράθυρα *Hanning* ορίζονται ως:

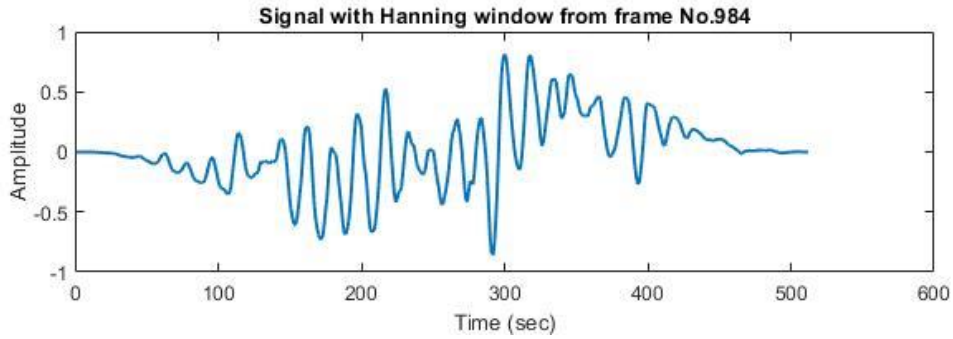
$$w(n) = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi n}{N}\right)\right]$$



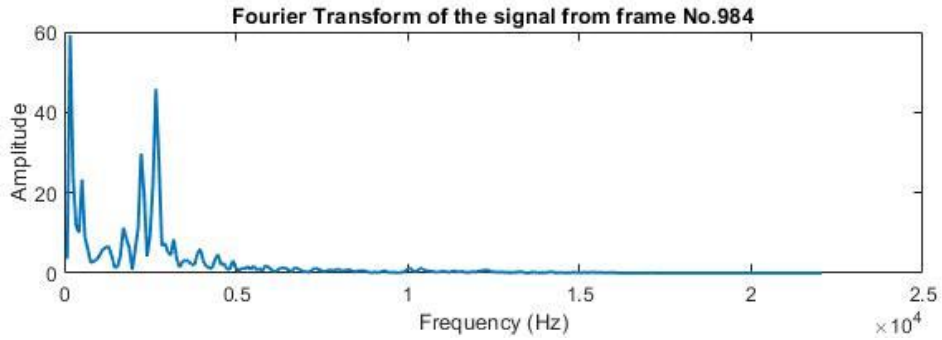
Εικόνα 4: Παράθυρο Hanning μήκους N=512 δειγμάτων.



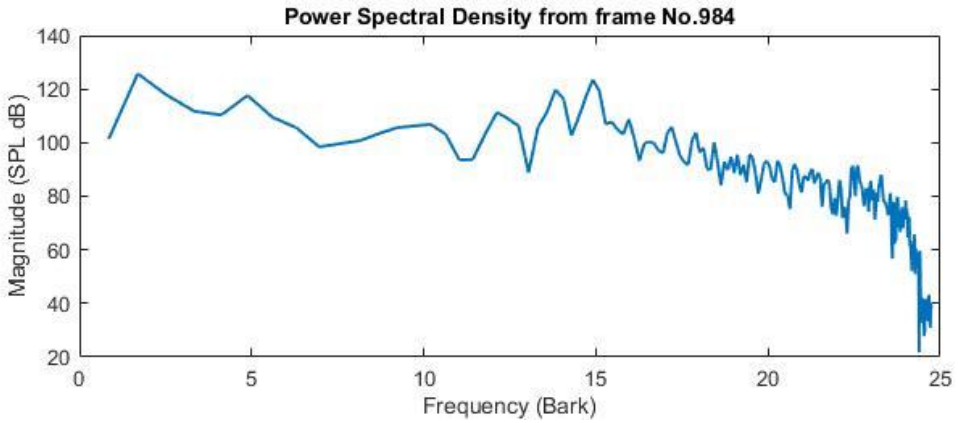
Εικόνα 5: Γραφική παράσταση του παραθύρου 984 του σήματος μουσικής σε συνάρτηση με το χρόνο.



Εικόνα 6: Γραφική παράσταση του παραθύρου 984 με τον χρόνο μετά από την εφαρμογή του παραθύρου Hanning.



Εικόνα 7: Μετασχηματισμός Fourier του σήματος.



Εικόνα 8: Φάσμα ισχύος του παραθύρου 984 του σήματος μουσικής.

Βήμα 1.2 : Εντοπισμός μασκών τόνων και θορύβου (Maskers)

Εφόσον έχει υπολογιστεί το φάσμα ισχύος *P(k)* κάθε παραθύρου συνεχίζουμε με την κυρίως επεξεργασία. Όσα αναφέρονται στη συνέχεια περιλαμβάνονται στην συνάρτηση *PsychoacousticModel()* και εφαρμόζονται σε κάθε πλαίσιο ξεχωριστά.

Η συνάρτηση λαμβάνει ως ορίσματα τον αριθμό του πλαισίου i, το φάσμα ισχύος του πλαισίου αυτού $P(i)$, τις συχνότητες στην κλίμακα *Bark* $b(f)$και το κατώφλι ακοής $T_q(f)$. Η έξοδος όπως έχει ήδη αναφερθεί είναι το συνολικό κατώφλι κάλυψης $T_g(i)$ για κάθε παράθυρο.

Στο τμήμα αυτό στόχος είναι ο εντοπισμός ανά *critical band* των τοπικών μεγίστων (μασκών), δηλαδή των συχνοτήτων στις οποίες η τιμή του φάσματος είναι μεγαλύτερη κατά τουλάχιστον 7dB από τις γειτονικές. Για τον εντοπισμό των

τονικών μασκών χρησιμοποιείται η συνάρτηση *findToneMaskers( )* η οποία δέχεται ως είσοδο το φάσμα ισχύος του πλαισίου i.

Η συνάρτηση *ST( )* που χρησιμοποιείται, επιστρέφει έναν πίνακα-γραμμή ο οποίος περιλαμβάνει λογικές τιμές {0,1} ανάλογα με το αν στην συχνότητα k υπάρχει μάσκα ή όχι. Η συνάρτηση αυτή εντοπίζει τις τονικές μάσκες εξετάζοντας αν υπάρχουν τοπικά μέγιστα σε διάφορες συχνοτικές περιοχές, σύμφωνα με τη σχέση:

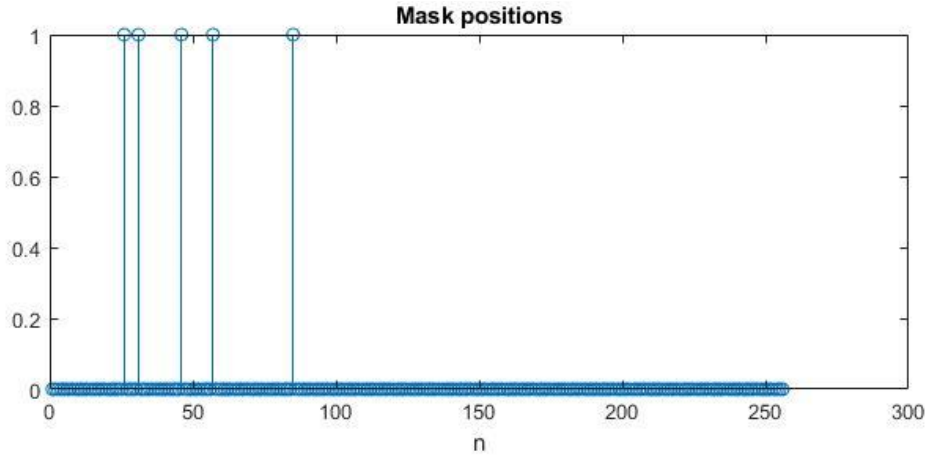
$$S_T(k) = \begin{cases} 0, & \alpha\nu\ k \notin [3,250] \\ P(k) > P(k \pm 1) \wedge P(k) > P(k \pm \Delta_k) + 7dB, & \alpha\nu\ k \in [3,250] \end{cases}$$

όπου

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

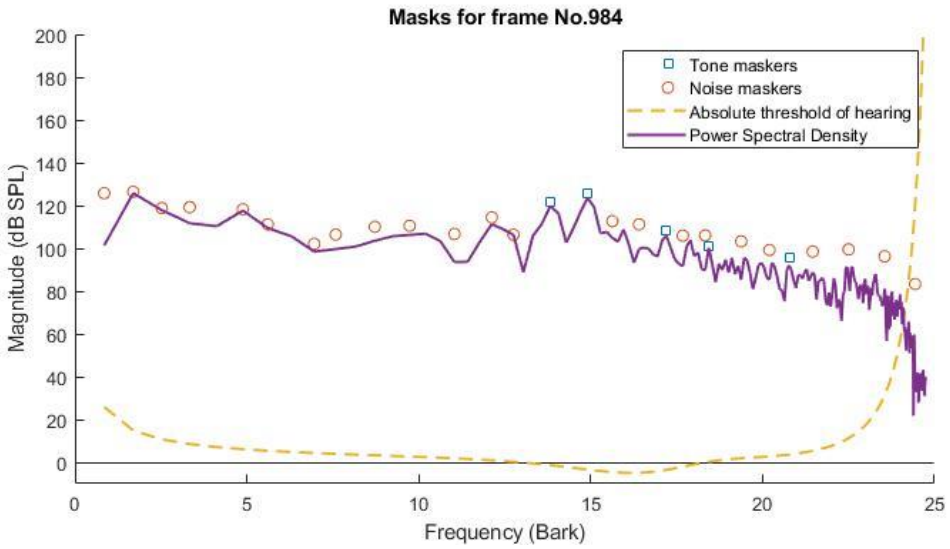
Εφόσον βρεθούν οι θέσεις των τονικών μασκών του πλαισίου, η συνάρτηση υπολογίζει την ισχύ τους, σύμφωνα με τη σχέση:

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



Εικόνα 9: Αποτελέσματα της συνάρτησης ST( ) για το εκατοστό παράθυρο του σήματος μουσικής.

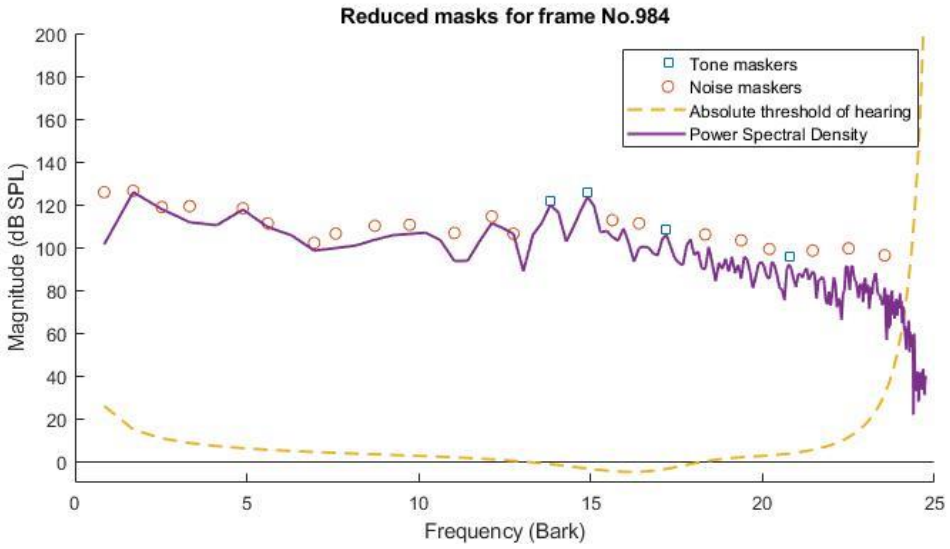
Για την εύρεση των μασκών θορύβου χρησιμοποιείται η έτοιμη συνάρτηση *findNoiseMaskers( )* που περιλαμβάνεται στα δεδομένα της άσκησης.



Εικόνα 10: Αποτελέσματα της συνάρτησης findToneMaskers( ). Στο διάγραμμα φαίνονται οι τονικές μάσκες και οι μάσκες θορύβου που βρέθηκαν, καθώς επίσης και το φάσμα ισχύος του παραθύρου.

Βήμα 1.3 : Μείωση και αναδιοργάνωση των μασκών

Στη συνέχεια είναι επιθυμητή η μείωση του αριθμού των μασκών, χρησιμοποιώντας την συνάρτηση *checkMaskers( )* που περιλαμβάνεται στο υλικό της άσκησης. Έξοδος της συνάρτησης είναι τα νέα διανύσματα P_{TM} και P_{NM}. Μέσω της συνάρτησης αυτής, απορρίπτεται κάθε μάσκα τόνου ή θορύβου που βρίσκεται κάτω από το κατώφλι ακοής *(Absolute Threshold of Hearing)*, ενώ στη συνέχεια χρησιμοποιώντας κινούμενα παράθυρα των 0.5 *Bark* διατηρούνται μόνο οι μάσκες με την μεγαλύτερη ένταση ανά περιοχή. Στα διαγράμματα που ακολουθούν φαίνονται οι μάσκες τόνου ή θορύβου που διατηρήθηκαν μετά από την παραπάνω επεξεργασία.



Εικόνα 11: Τονικές μάσκες και μάσκες θορύβου του πλαισίου 984 του σήματος μουσικής μετά από τη διαδικασία μείωσης και αναδιοργάνωσης των μασκών. Φαίνονται επίσης το Συνολικό Κατώφλι Ακοής και το Φάσμα Ισχύος του σήματος.

Βήμα 1.4 : Υπολογισμός των δύο διαφορετικών κατωφλίων κάλυψης *(Individual Masking Thresholds)*

Εφόσον έχουν βρεθεί τα σωστά διανύσματα των μασκών θορύβου και τόνων, υπολογίζονται τα δύο διαφορετικά κατώφλια κάλυψης:

$$T_{TM}(i,j) = P_{TM}(j) - 0.275b(j) + SF(i,j) - 6.025(dBSPL)$$

$$T_{NM}(i,j) = P_{NM}(j) - 0.175b(j) + SF(i,j) - 2.025(dBSPL)$$

Τα κατώφλια που προκύπτουν, αντιπροσωπεύουν το ποσοστό κάλυψης στο σημείο i, από τη μάσκα θορύβου ή τόνου που βρίσκεται στο σημείο j. Συνεπώς οι πίνακες T_{TM} και T_{NM} που προκύπτουν έχουν διαστάσεις $\binom{N}{2} \times A$ όπου A είναι ο αριθμός των μασκών τόνων ή θορύβου αντίστοιχα, οι οποίες βρέθηκαν στο πλαίσιο που επεξεργαζόμαστε. Στις περιοχές όπου $b(i) \notin [b(j) - 3, b(j) + 8]$ οι τιμές των T_{TM} και T_{NM} είναι μηδενικές εφόσον θεωρούμε, σύμφωνα με το μοντέλο, ότι εκεί δεν υπάρχει κάλυψη από την μάσκα j. Η συνάρτηση *SF(i,j)* προσδιορίζει την ελάχιστη τιμή ισχύος που πρέπει να έχουν οι γειτονικές συχνότητες ώστε να γίνουν αντιληπτές από τον άνθρωπο.

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

όπου $\Delta_b = b(i) - b(j)$ η διαφορά συχνοτήτων μεταξύ των θέσεων j (της μάσκας) και i (του σημείου της γειτονιάς που θεωρούμε πως υπάρχει κάλυψη), σε κλίμακα *Bark*.

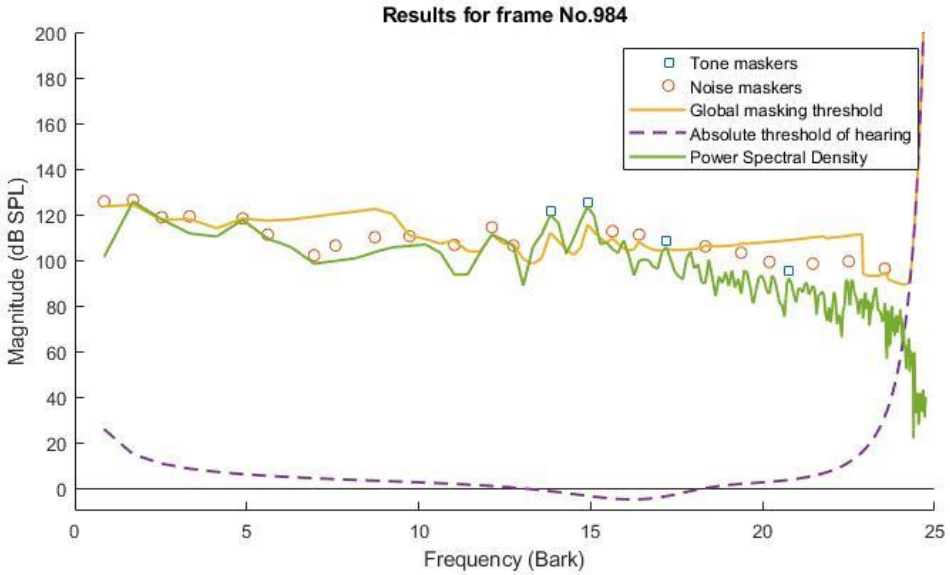
Για τον υπολογισμό του κατωφλίου κάλυψης των μασκών θορύβου αντικαθιστούμε στην παραπάνω σχέση το $P_{TM}(j)$ με $P_{NM}(j)$.

Βήμα 1.5 : Υπολογισμός του συνολικού κατωφλίου κάλυψης *(Global Masking Threshold)*

Τέλος, για κάθε πλαίσιο υπολογίζουμε το συνολικό κατώφλι κάλυψης σε κάθε μία από τις διακριτές συχνότητες ξεχωριστά. Το συνολικό κατώφλι υπολογίζεται σύμφωνα με την σχέση:

$$T_g(k) = 10\log_{10}\left(10^{0.1T_q(k)} + \sum_{l=1}^{L} 10^{0.1T_{TM}(k,l)} + \sum_{m=1}^{M} 10^{0.1T_{NM}(k,m)}\right) \; dB\ SLP$$

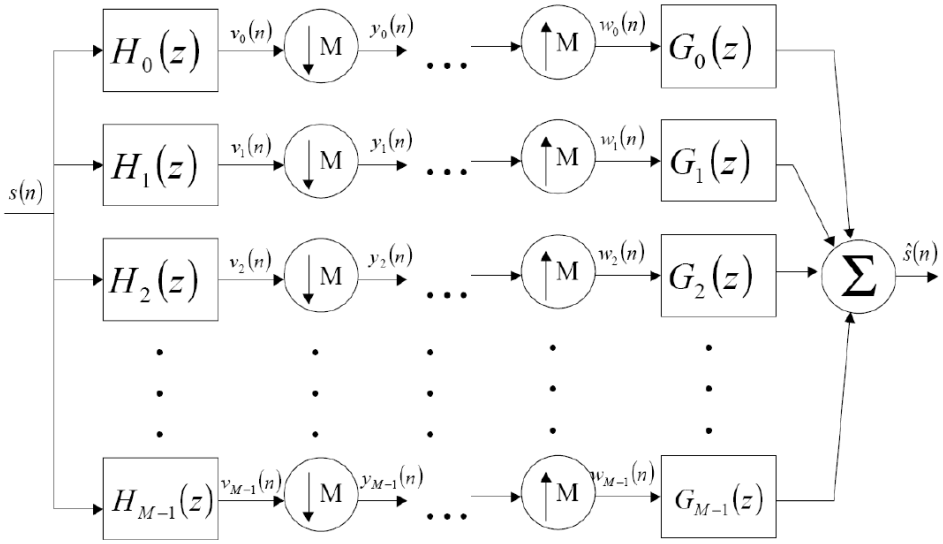
όπου το $T_q(k)$ είναι το *Absolute Threshold of Hearing* σε κάθε διακριτή συχνότητα k, $T_{TM}(k,l)$ και $T_{NM}(k,m)$ τα κατώφλια κάλυψης που υπολογίστηκαν προηγουμένως και L, M ο αριθμός των μασκών τόνου και θορύβου αντίστοιχα.



Εικόνα 12: : Αποτελέσματα της συνάρτησης P*sychoacousticModel( ) για το παράθυρο 984 του σήματος μουσικής.*

Μέρος 2. Χρονο-Συχνοτική ανάλυση με συστοιχία ζωνοπερατών φίλτρων

Στόχος του δεύτερου μέρους είναι, όπως ήδη αναφέρθηκε, η δημιουργία μιας συνάρτησης η οποία μέσω της διαδικασίας που φαίνεται στο *Σχήμα 1* κωδικοποιεί το σήμα μουσικής. Η χρήση συστοιχίας φίλτρων, μέσω της διαίρεσης του φάσματος σε ζώνες συχνοτήτων και συνεπώς της εξαγωγής χαρακτηριστικών για την συχνοτική κατανομή του συνολικού σήματος, βοηθά στην ταυτοποίηση των αντιληπτικά περιττών σημείων και άρα στη μείωση στατιστικών λαθών στο τελικό κωδικοποιημένο σήμα.



Σχήμα 1: Uniform M-Band Maximally Decimated Analysis-Synthesis Filterbank.

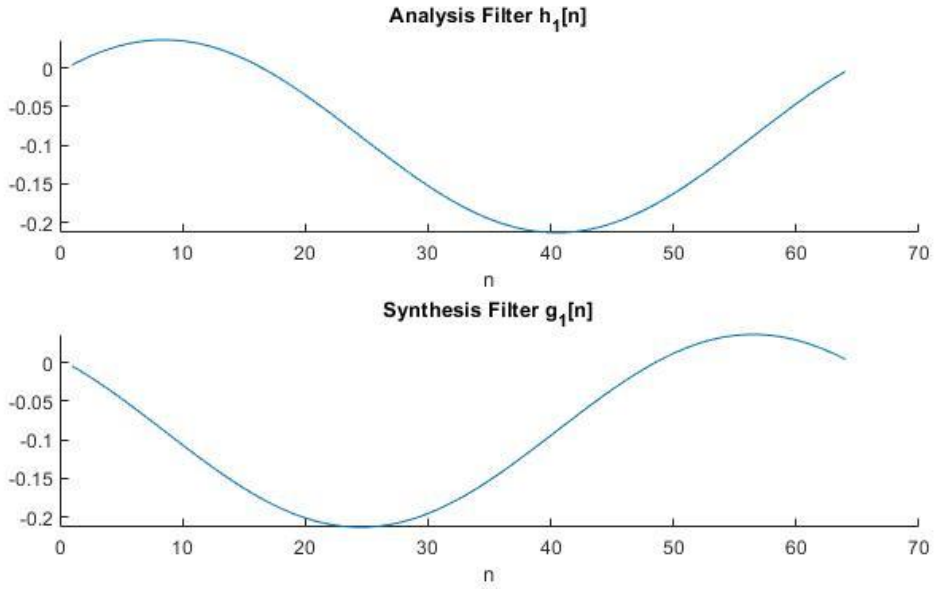
Βήμα 2.0 : Συστοιχία ζωνοπερατών φίλτρων

Αρχικά δημιουργούμε τις συστοιχίες ζωνοπερατών φίλτρων ανάλυσης και σύνθεσης χρησιμοποιώντας τον *Modified Discrete Cosine Transform (MDCT*). Κάθε συστοιχία περιλαμβάνει *M=32* φίλτρα, οι αντίστοιχες κρουστικές αποκρίσεις των οποίων είναι:

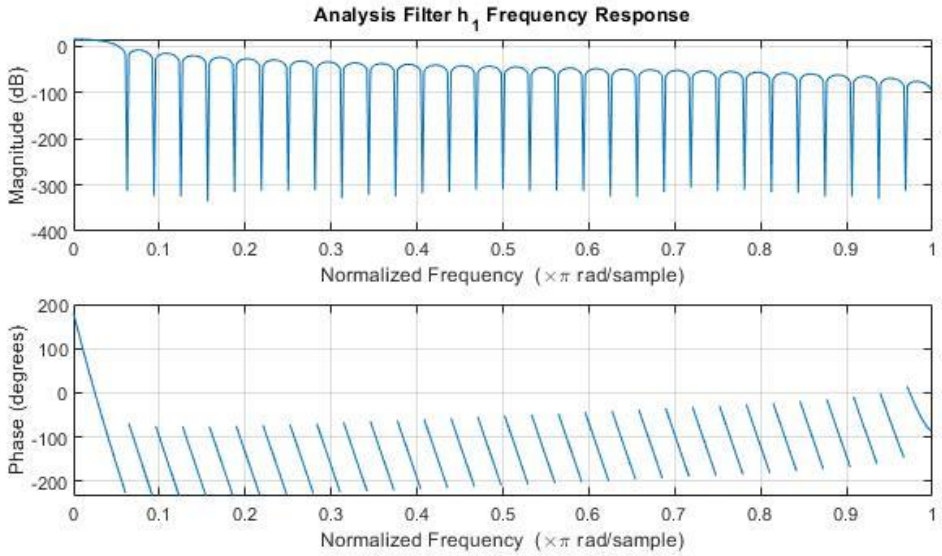
$$h_k(n) = \sin\left[\left(n+\frac{1}{2}\right)\frac{\pi}{2M}\right]\sqrt{\frac{2}{M}}\cos\left[\frac{(2n+M+1)\cdot(2k+1)\cdot\pi}{4M}\right]$$

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

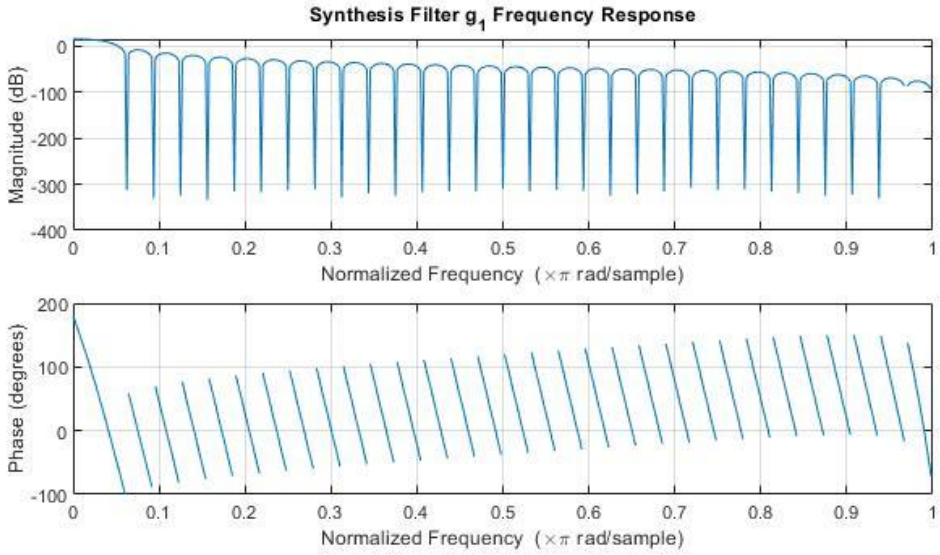
Όπου $L = 2M$, $0 \le n \le L-1$ και $0 \le k \le M-1$.



Εικόνα 13: Απόκριση του πρώτου φίλτρου των συστοιχιών ανάλυσης και σύνθεσης.



Εικόνα 14: Συχνοτική απόκριση του πρώτου φίλτρου ανάλυσης.



Εικόνα 15: Συχνοτική απόκριση του πρώτου φίλτρου σύνθεσης.

Βήμα 2.1 : Ανάλυση με συστοιχία φίλτρων

Στη συνέχεια μέσω συνέλιξης κάθε παραθυροποιημένου σήματος μουσικής με τα 32 φίλτρα ανάλυσης προκύπτει το σήμα:

$$u_k(n) = h_k(n) * x(n)$$

Το σήμα αυτό υποδειγματοληπτείται κατά παράγοντα M, χωρίς επικαλύψεις, λαμβάνοντας τελικά:

$$y_k(n) = u_k(Mn)$$

Βήμα 2.2 : Κβαντοποίηση

Στη συνάρτηση που υλοποιήθηκε γίνεται χρήση δύο διαφορετικών κβαντιστών. Αρχικά ένας προσαρμοζόμενος ομοιόμορφος κβαντιστής 2^{B_k} επιπέδων, όπου B_k είναι ο αριθμός των *bits* κωδικοποίησης ανά δείγμα της ακολουθίας $y_k(n)$ στο τρέχον πλαίσιο του σήματος μουσικής. Το βήμα του κβαντιστή αυτού μεταβάλλεται σε κάθε πλαίσιο ανάλυσης με βάση τη σχέση:

$$B_k = \log_2\left(\frac{2^{16}}{\min(T_g(i))}\right) - 1$$

Όπου το αρχικό σήμα έχει κωδικοποιηθεί με *16bits* επομένως έχει $2^{16} = 65536$ διακριτές βαθμίδες έντασης. Επίσης τα επίπεδα του κβαντιστή προσαρμόζονται κάθε φορά με βάση τα όρια του πλαισίου ανάλυσης, ώστε να επιτευχθεί μείωση του σφάλματος.

Η δεύτερη μέθοδος που χρησιμοποιείται, με σκοπό τη σύγκριση των αποτελεσμάτων, είναι ένας μη-προσαρμοζόμενος κβαντιστής των *8bits*, με σταθερό βήμα κβαντισμού και θεωρώντας ότι τα όριά του είναι [-1,1].

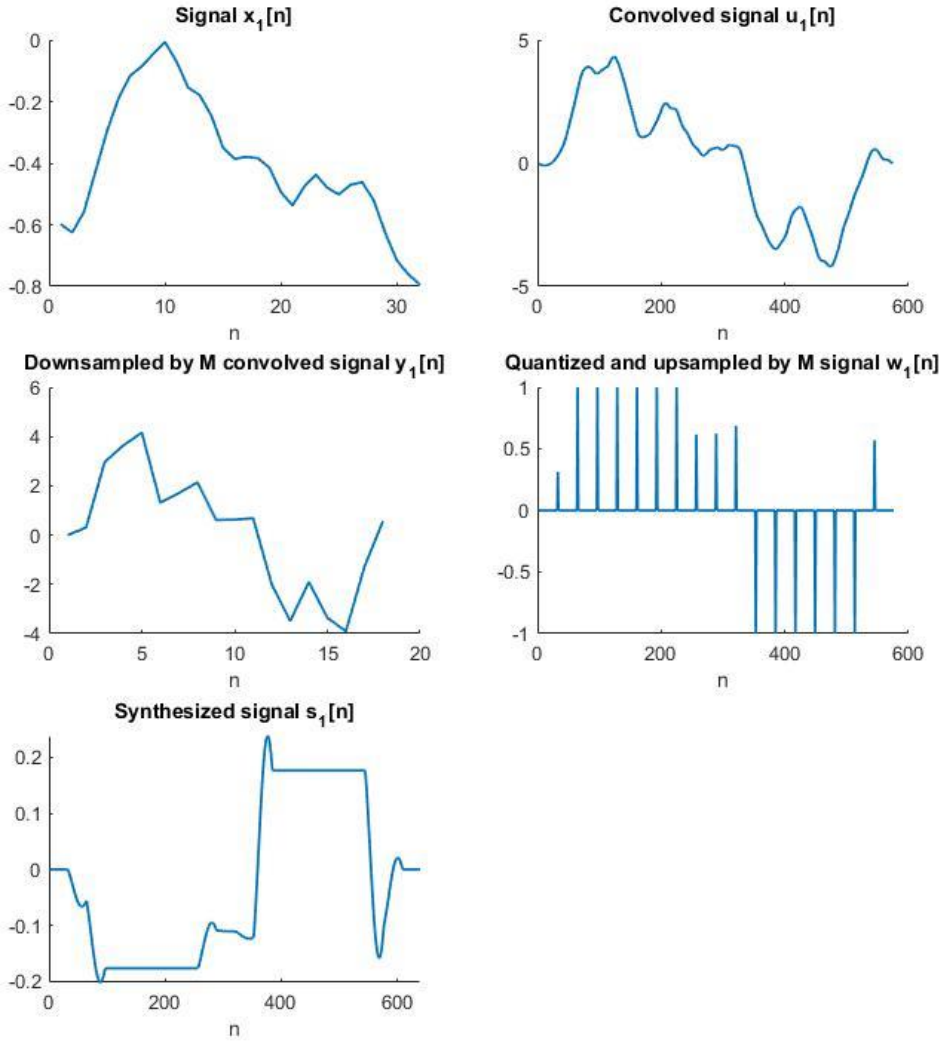
Βήμα 2.3 : Σύνθεση

Οι κβαντισμένες ακολουθίες $\tilde{y}_k(n)$ περνούν από τον αποκωδικοποιητή όπου υπερδειγματοληπτούνται κατά παράγοντα M, οπότε προκύπτει:

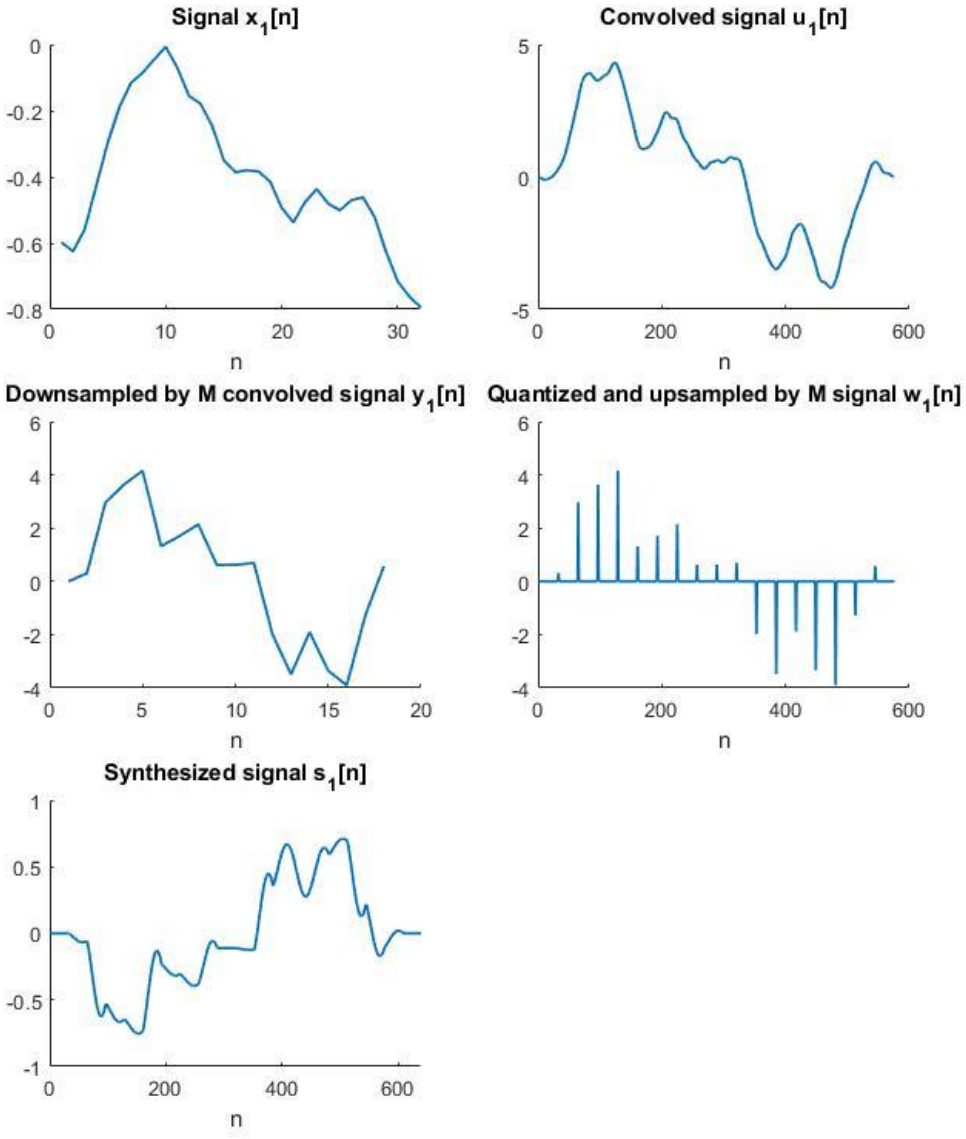
$$w_k(n) = \begin{cases} \tilde{y}_k\left({}^{n}/_{M}\right), & n = 0, M, 2M, \dots \\ 0\,, & \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$$

Τέλος μέσω συνέλιξης με τα φίλτρα σύνθεσης λαμβάνουμε το ανακατασκευασμένο σήμα:

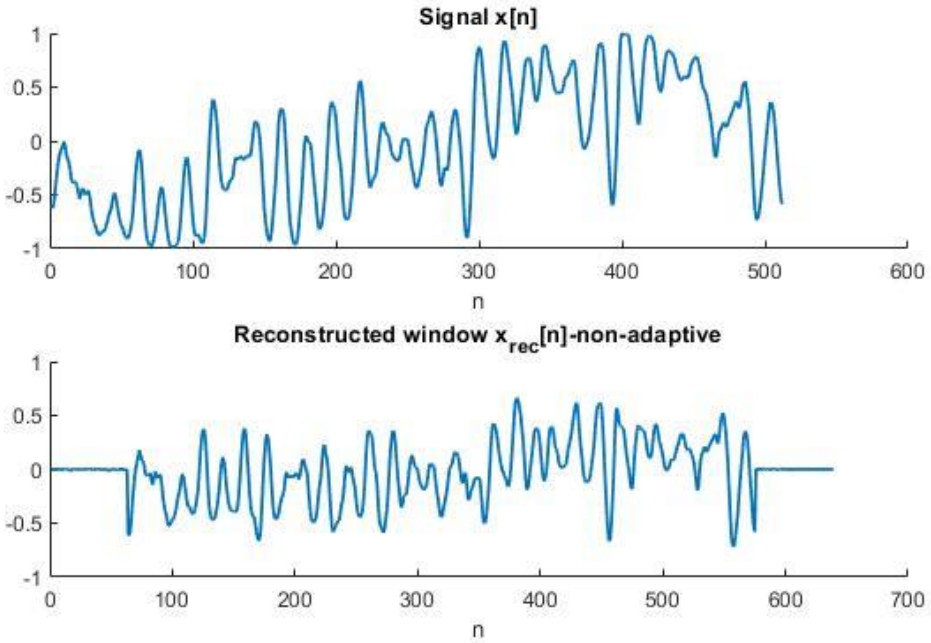
$$y_{rec}(n) = g_k(n) * w_k(n)$$



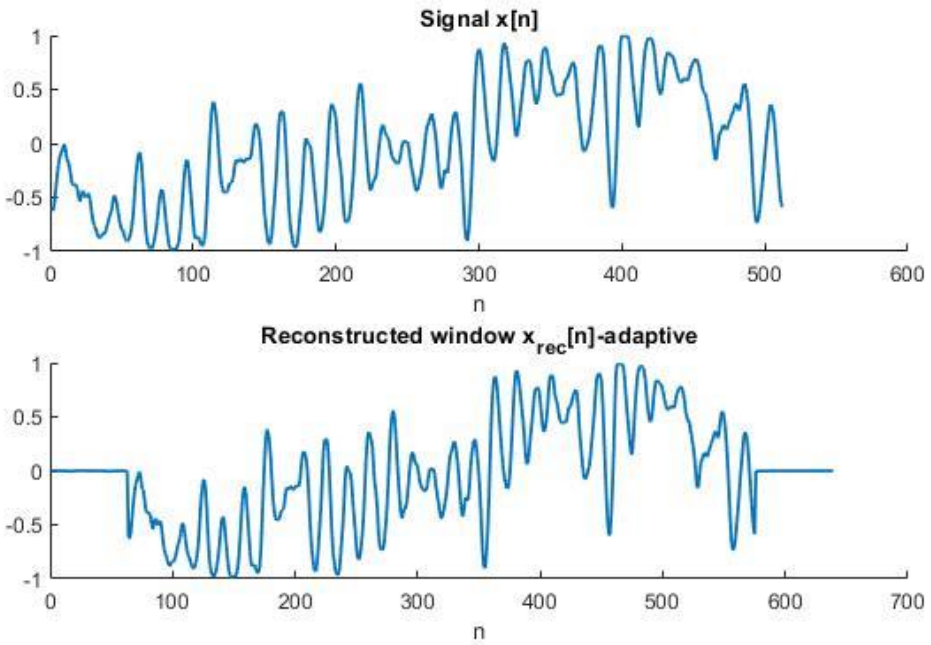
Εικόνα 16: Κωδικοποίηση του παραθύρου 984 με χρήση του μη-προσαρμοζόμενου κβαντιστή των 8 bits.



Εικόνα 17: Κωδικοποίηση του παραθύρου 984 με χρήση του προσαρμοζόμενου κβαντιστή.



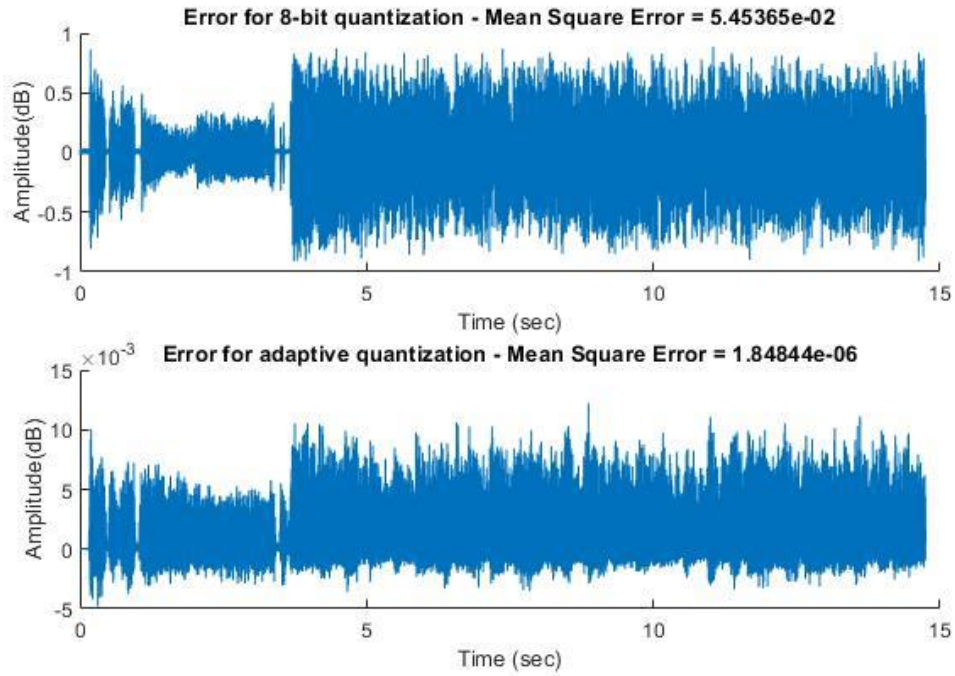
Εικόνα 18: Αρχικό και ανακατασκευασμένο σήμα του επιλεγμένου πλαισίου με χρήση του μη-προσαρμοζόμενου κβαντιστή των 8bits.



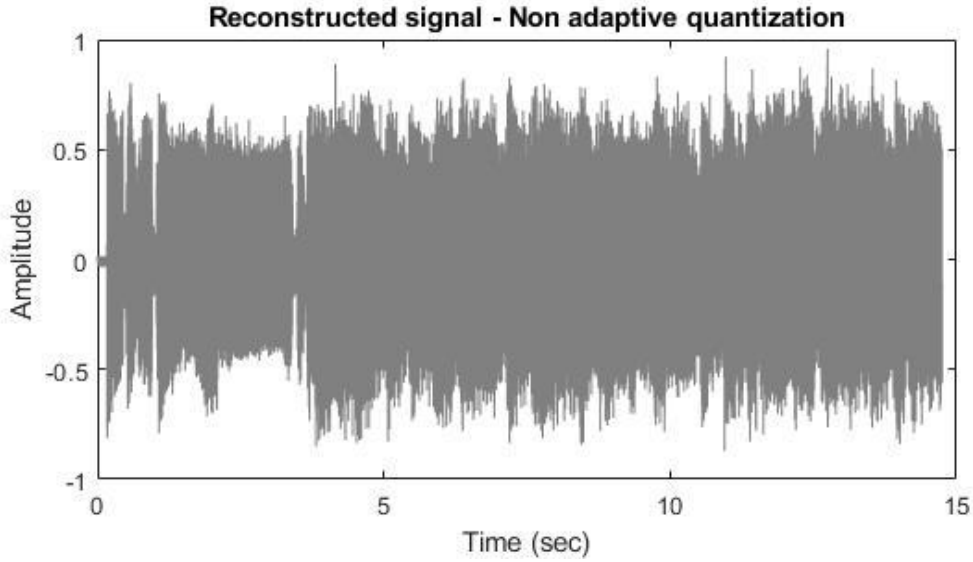
Εικόνα 19: Αρχικό και ανακατασκευασμένο σήμα του επιλεγμένου πλαισίου με χρήση του προσαρμοζόμενου κβαντιστή.

Το τελικό ανακατασκευασμένο σήμα μουσικής προκύπτει με χρήση της τεχνικής *OverLap-Add* χωρίς επικάλυψη μεταξύ των διαδοχικών παραθύρων ανάλυσης. Η έξοδος της συστοιχίας των φίλτρων είναι μετατοπισμένη κατά 63 δείγματα σε σχέση με την είσοδο, γεγονός που λαμβάνεται υπόψιν στο τελικό σήμα ανακατασκευής. Επίσης αφαιρούνται τα μηδενικά δείγματα στο τέλος του σήματος.

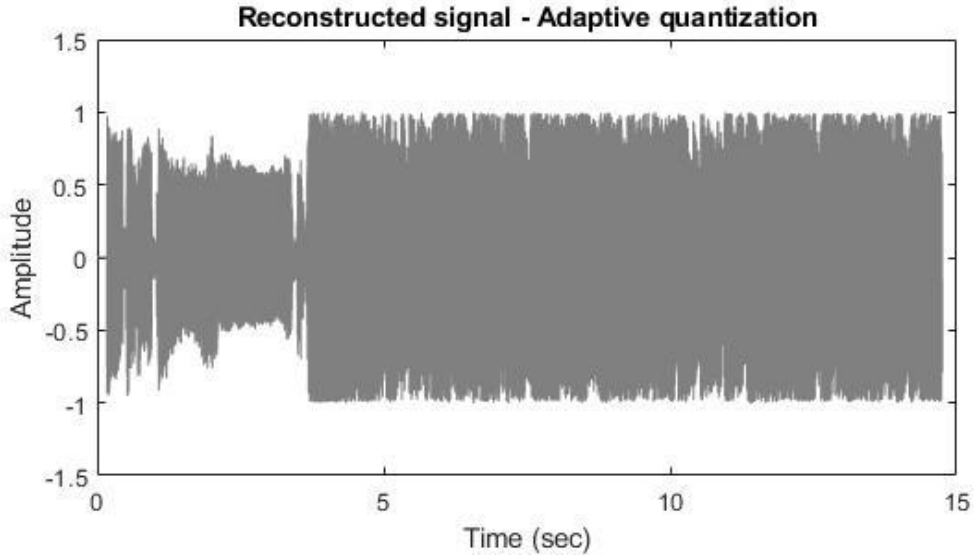
Τα αποτελέσματα φαίνονται στη συνέχεια.



Εικόνα 20: Σφάλμα των ανακατασκευασμένων σημάτων σε σχέση με το αρχικό, σε συνάρτηση με τον χρόνο.



Εικόνα 21: Ανακατασκευασμένο σήμα με χρήση μη- προσαρμοζόμενου κβαντιστή.



Εικόνα 22: Ανακατασκευασμένο σήμα με χρήση προσαρμοζόμενου κβαντιστή.

Συγκρίνοντας τα δύο αποτελέσματα προκύπτουν τα εξής για τα ανακατασκευασμένα σήματα μουσικής:

- Μη-προσαρμοζόμενος κβαντιστής 8bits:
  - Μέσο τετραγωνικό σφάλμα: 0.054537
  - Ποσοστό συμπίεσης: 50%
- Προσαρμοζόμενος κβαντιστής:
  - Μέσο τετραγωνικό σφάλμα: $1.848438 \cdot 10^{-6}$
  - Ποσοστό συμπίεσης: 0.1105%