Ε.Μ.Π., Σχολή Η.Μ.& Μ.Υ. **Ψηφιακή Επεξεργασία Σημάτων** Ακαδ.Ετος 2014-**2015**
Ημερ/νια: 24/04/2015 **2η Εργαστηριακή Ασκηση** Παραδοτέα: 15/05/2015

[Εξηγείστε περιεκτικά και επαρκώς την εργασία σας. Επιτρέπεται προαιρετικά η συνεργασία εντός ομάδων των 2 ατόμων. Κάθε ομάδα 2 ατόμων υποβάλλει μια κοινή αναφορά που αντιπροσωπεύει μόνο την προσωπική εργασία των μελών της. Αν χρησιμοποιήσετε κάποια άλλη πηγή εκτός του βιβλίου και του εκπαιδευτικού υλικού του μαθήματος, πρέπει να το αναφέρετε. Η παράδοση της αναφοράς και του κώδικα της εργασίας θα γίνει ηλεκτρονικά στο mycourses.ntua.gr και επιπλέον η αναφορά της εργασίας θα παραδίδεται τυπωμένη και προσωπικά στην γραμματεία του εργαστηρίου Ρομποτικής (2.1.12, παλαιό Κτ.Ηλεκ.), ώρες 09.00-14.00].

Γραμμική Πρόβλεψη (LPC) και Ομομορφική (Cepstrum) Επεξεργασία Σημάτων με MATLAB και Εφαρμογές στη Σύνθεση και Συμπίεση Φωνής

Μέρος 1. Εξαγωγή pitch φωνής με τη χρήση Cepstrum

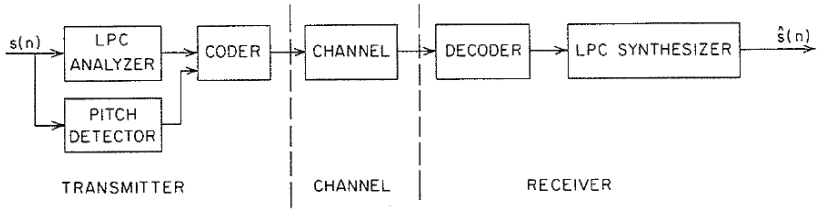
Στόχος του πρώτου μέρους της εργασίας είναι να υλοποιηθεί αλγόριθμος εξαγωγής της θεμελιώδους συχνότητας της φωνής με τη χρήση ομομορφικής επεξεργασίας σημάτων. Θα πρέπει να ακολουθήσετε τις οδηγίες που δίνονται στην Ενότητα 7.3 του βιβλίου των Rabiner & Schafer, 1978. Σας δίνεται το αρχείο φωνής όπου θα πρέπει να εφαρμόσετε τον αλγόριθμό σας speech.wav καθώς επίσης και ένα αρχείο pitch.mat (βλ. Παράρτημα Β) που περιέχει τη θεμελιώδη συχνότητα όπως έχει εξαχθεί με έναν αλγόριθμο που βασίζεται στην αυτοσυσχέτιση. Θα πρέπει να υπολογίσετε τη διαφορά μεταξύ της μεταβαλλόμενης με το χρόνο συχνότητας που υπολογίζετε εσείς για το σήμα φωνής και της συχνότητας που σας δίνεται. Για άφωνους ήχους θεωρήστε ότι η θεμελιώδης συχνότητα είναι μηδέν. Αναπαραστήστε και γραφικά τις δύο χρονοσειρές συχνοτήτων για να είναι ευκολότερη η σύγκρισή τους.

Μέρος 2. Ψηφιακή Σύνθεση Φωνής με Γραμμική Πρόβλεψη (LPC Vocoder)

Μια από τις πρώτες και πολύ βασικές εφαρμογές της γραμμικής πρόβλεψης είναι η κωδικοποίηση/συμπίεση φωνής για μετάδοσή της και αποθήκευση. Δύο σημαντικά μέρη αυτού του συστήματος, όπως φαίνεται στο διάγραμμα του Σχήματος 1, είναι η ανάλυση του σήματος φωνής για την κωδικοποίηση του και η ανασύνθεσή του μετά την αποκωδικοποίηση. Η παρούσα εργασία αποσκοπεί στο σχεδιασμό αυτών των δύο συστημάτων για την ψηφιακή σύνθεση φωνής. Για την ανάλυση θα χρησιμοποιηθεί γραμμική πρόβλεψη, ενώ για την ανασύνθεση θα χρησιμοποιηθεί το αντίστροφο σύστημα γραμμικής πρόβλεψης με τη μέθοδο Overlap-Add.

2.1 Ανάλυση Φωνής με Γραμμική Πρόβλεψη

Υλοποιήστε σύστημα LPC μοντελοποίησης σημάτων χρησιμοποιώντας την μέθοδο autocorrelation με Hamming παράθυρο και τάξη προβλέπτη p. Το σύστημα θα πρέπει να δέχεται στην είσοδο ένα πλαίσιο ανάλυσης (analysis frame) του σήματος $s[n]$, να το παραθυρώνει



Σχήμα 1: Κωδικοποιητής Φωνής: LPC Vocoder.

κατάλληλα και να δημιουργεί το σήμα $x[n]$, να εφαρμόζει τη μέθοδο autocorrelation στο $x[n]$, και να επιστρέφει το μοντέλο γραμμικής πρόβλεψης που προκύπτει:

$$H(z) = \frac{G}{A(z)}, \quad A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k}$$

δηλαδή τους αντίστοιχους συντελεστές γραμμικής πρόβλεψης α_k, το κέρδος $G = \sqrt{R[0] - \sum_{k=1}^{p} \alpha_k R[k]}$, και το σήμα λάθους της γραμμικής πρόβλεψης

$$e(n) = x[n] - \sum_{k=1}^{p} \alpha_k x[n-k]$$

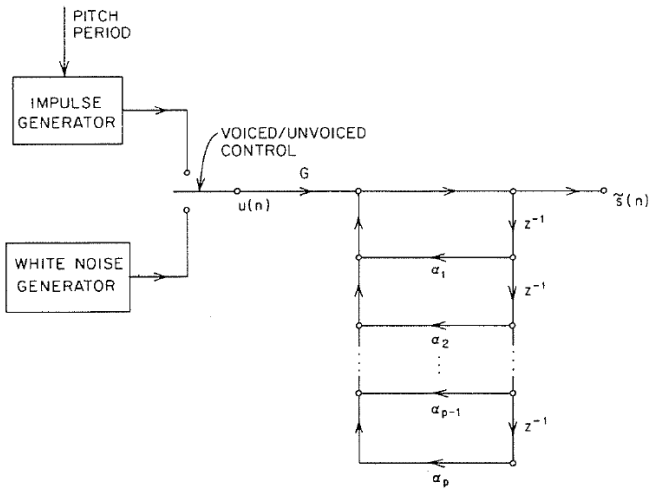
Για ανάλυση φωνής επιλέγεται συνήθως η τιμή $p \approx F_s(kHz) + 4$ ως ικανοποιητική τάξη προβλέπτη για μοντελοποίηση της φασματικής πολυπλοκότητας. Για κάθε σήμα $s[n]$ εφαρμόστε το σύστημα LPC μοντελοποίησης και σε κοινό διάγραμμα αναπαραστήστε γραφικά το φάσμα του λάθους της γραμμικής πρόβλεψης $e(n)$ για δύο τιμές της τάξης του μοντέλου p. Είναι το λάθος πρόβλεψης λευκός θόρυβος; Για τις ίδιες τιμές της τάξης του μοντέλου αναπαραστήστε γραφικά σε κοινό διάγραμμα το φάσμα του LPC μοντέλου ($20 \log_{10} |H(e^{j\omega})|$) και το φάσμα του παραθυρωμένου σήματος $x[n]$ όπως υπολογίζεται με FFT. (Σημείωση: Το LPC φάσμα μπορείτε να το υπολογίσετε εύκολα και με FFT.)
Χρήσιμες MATLAB υπορουτίνες: **fft, hamming, lpc, filter, freqz**

2.2 Σύνθεση Φωνής με Γραμμική Πρόβλεψη

Δεδομένου του ότι ένα σήμα φωνής $s[n]$ δεν μπορεί να θεωρηθεί παρά τοπικά μόνο στάσιμο, η ανάλυση του με γραμμική πρόβλεψη είναι αναγκαίο να γίνει σε μικρά επικαλυπτόμενα πλαίσια ανάλυσης. Η επικάλυψη είναι αναγκαία ώστε να διασφαλιστεί η συνέχεια και η ομαλότητα της ανάλυσης. Θεωρήστε ότι κάθε πλαίσιο ανάλυσης του παραθυροποιημένου σήματος $x[n]$ έχει μήκος 30 msec και η επικάλυψή του με το προηγούμενο πλαίσιο είναι ίση με $2L/3$ όπου L το μήκος του πλαισίου ανάλυσης. Υλοποιήστε σύστημα που θα δέχεται στην είσοδο το σήμα φωνής $s[n]$, θα το χωρίζει σε επιμέρους πλαίσια ανάλυσης και θα επιστρέφει για το καθένα το αντίστοιχο μοντέλο γραμμικής πρόβλεψης, όπως περιγράφεται στο μέρος 1.2 της εργασίας.

Δεδομένα: Σήμα φωνής speech.wav.

Η φωνή μπορεί να ανασυντεθεί από τις LPC παραμέτρους χρησιμοποιώντας το σύστημα που φαίνεται στο Σχήμα 2. Εκτός από τις παραμέτρους της γραμμικής πρόβλεψης, ο συνθέτης χρειάζεται και τη θεμελιώδη περίοδο καθώς και την πληροφορία αν ο ήχος που συντίθεται είναι έμφωνος ή άφωνος. Με αυτόν τον τρόπο είναι δυνατός ο κατάλληλος προσδιορισμός του σήματος $u[n]$ που διεγείρει το μοντέλο γραμμικής πρόβλεψης. Το σήμα $u[n]$ είναι είτε



Σχήμα 2: LPC Συνθέτης Φωνής.

μια περιοδική παλμοσειρά με τη δεδομένη θεμελιώδη περίοδο (pitch) για έμφωνους ήχους είτε μια ακολουθία ομοιόμορφα κατανεμημένων τυχαίων δειγμάτων με μηδενική μέση τιμή και μοναδιαία τυπική απόκλιση, για άφωνους ήχους. Για ευκολία, θεωρήστε ότι έχετε άφωνο ήχο όταν η τιμή της θεμελιώδους περιόδου είναι μηδενική. Σημειώστε ότι, όπως αναφέρεται και στην ανάλυση στην ενότητα 8.2 του βιβλίου των Rabiner & Schafer, σύμφωνα με το χρησιμοποιούμενο μοντέλο, η διέγερση $Gu[n]$ θα πρέπει να έχει την ίδια ενέργεια με το λάθος πρόβλεψης $e[n]$.
Αν στην είσοδο δίνονται σύνολα παραμέτρων που έχουν προκύψει από διαδοχικά επικαλυπτόμενα πλαίσια ανάλυσης τότε το συντεθειμένο σήμα για κάθε πλαίσιο ανάλυσης θα πρέπει να συνδυαστεί κατάλληλα με τα γειτονικά του ώστε η σύνθεση του συνολικού σήματος φωνής να είναι επιτυχής. Για το σκοπό αυτό χρησιμοποιείται η τεχνική Overlap-Add (βλ. Παράρτημα Α), που θεωρεί πως τα πλαίσια συνθετικής φωνής εμφανίζουν την ίδια επικάλυψη με αυτή που θεωρήθηκε κατά την ανάλυση, οπότε το τελικό σήμα φωνής (με ακρίβεια μιας πολλαπλασιαστικής σταθεράς) προκύπτει με την κατάλληλη (αφού το κάθε πλαίσιο τοποθετηθεί σωστά στο χρόνο) πρόσθεση των επικαλυπτόμενων πλαισίων.
Χρήσιμες MATLAB υπορουτίνες: **buffer, fft, hamming, lpc, filter, rand**

ΠΑΡΑΔΟΤΕΑ: Φασματογράφημα (Spectrogram) του αρχικού σήματος φωνής και του συνθετικού, καθώς και ένα αρχείο synthesis_lpc.wav με την συνθετική φωνή.

Μέρος 3. Βελτιωμένη Σύνθεση Φωνής με Χρήση Μακροπρόθεσμης Πρόβλεψης και Κατάλληλα Σχεδιασμένης Βάσης Διεγέρσεων (CELP)

3.1 Εισαγωγή προβλέπτη μακράς καθυστέρησης

Στόχος είναι να υλοποιηθεί ένας προβλέπτης που θα προβλέπει την περιοδικότητα της φωνής όπου υπάρχει και θα ακολουθεί τον πρώτο προβλέπτη στην ανάλυση. Διαδοχικές περίοδοι στα τμήματα έμφωνων ήχων παρουσιάζουν αρκετή ομοιότητα. Η ίδια περιοδικότητα περίπου εμφανίζεται και στο λάθος πρόβλεψης $e(n)$ του πρώτου προβλέπτη η οποία μπορεί να αφαιρεθεί εφαρμόζοντας έναν νέο προβλέπτη 3ης τάξης $P_e = \beta_1 z^{-M+1} + \beta_2 z^{-M} + \beta_3 z^{-M-1}$ όπου M η μακρά καθυστέρηση που αντιστοιχεί σε ακέραια πολλαπλάσια της περιόδου[1] του $e[n]$ και $\beta_1, \beta_2, \beta_3$ οι συντελεστές γραμμικής πρόβλεψης που μπορούν να βρεθούν ελαχιστο-

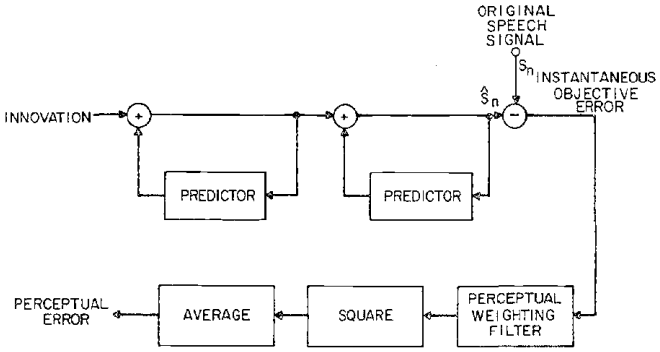
[1] συνήθως από $2ms$ έως $20ms$ για σήματα με συχνότητα δειγματοληψίας $8kHz$

ποιώντας το τετράγωνο του λάθους της νέας πρόβλεψης $\hat{e}[n]$. Η ελαχιστοποίηση οδηγεί σ'ενα σύστημα γραμμικών εξισώσεων. α) Υπολογίστε το φίλτρο P_e για έναν έμφωνο ήχο. Δείξτε τα λάθη $e[n], \hat{e}[n]$ των δύο προβλεπτών σε αντιπαραβολή με το αρχικό σήμα. Είναι το λάθος $\hat{e}[n]$ λευκός θόρυβος; β) Εισάγετε κατάλληλα το φίλτρο στον συνθέτη που υλοποιήσατε στο 2.2 και σχολιάστε σχετικά με την ποιότητα της σύνθεσης συγκρίνοντας με το προηγούμενο σύστημα.

ΠΑΡΑΔΟΤΕΑ: Το φασματογράφημα (Spectrogram) της σύνθεσης και η ηχογράφιση της στο αρχείο synthesis_lpc_long.wav.

3.2 Δημιουργία βάσης τυχαίων διεγέρσεων και επιλογή της βέλτιστης για σύνθεση

Η ποιότητα της σύνθεσης φωνής μπορεί να βελτιωθεί αισθητά με τη χρησιμοποίηση καταλληλότερων διεγέρσεων. Η επιλογή της βέλτιστης διέγερσης από ένα συγκεκριμένο σύνολο μπορεί να πραγματοποιηθεί μέσω της ελαχιστοποίησης του λάθους της σύνθεσης που υπολογίζεται από το σύστημα που δίνεται στο Σχήμα 3. Στόχος είναι να υλοποιηθεί αυτός



Σχήμα 3: Ελαχιστοποίηση λάθους σύνθεσης

ο αλγόριθμος ελαχιστοποίησης ώστε να επιτρέπεται η επιλογή της καταλληλότερης διέγερσης από προκαθορισμένο σύνολο. Το σύνολο των διεγέρσεων θα περιλαμβάνει σήματα της μορφής:

$$v_n = \sum_{k=0}^{N-1} c_k \cos(\pi k n/N + \phi_k), n = 0, 1, \ldots, 2N-1 \quad (1)$$

όπου c_k και ϕ_k είναι ανεξάρτητες τυχαίες μεταβλητές. Η ϕ_k είναι ομοιόμορφα κατανεμημένη μεταξύ 0 και 2π και η c_k είναι Rayleigh-κατανεμημένη με συνάρτηση πυκνότητας πιθανότητας:

$$p(c_k) = c_k \exp(-c_k^2/2), c_k > 0. \quad (2)$$

Θεωρείστε οτι έχετε διαθέσιμα 10 bits για να κωδικοποιείσετε την πληροφορία της διέγερσης ανα πλαίσιο σύνθεσης. Συνεπώς, η βάση των διεγέρσεων που θα δημιουργείσετε θα περιέχει 1024 τυχαία σήματα v_n. Εισάγετε κατάλληλα το σύστημα που υλοποιήσατε στο συνθέτη φωνής και σχολιάστε σχετικά με την ποιότητα της συνθετικής φωνής.

ΠΑΡΑΔΟΤΕΑ: Το φασματογράφημα (Spectrogram) της σύνθεσης και η ηχογράφιση της στο αρχείο synthesis_celp.wav.

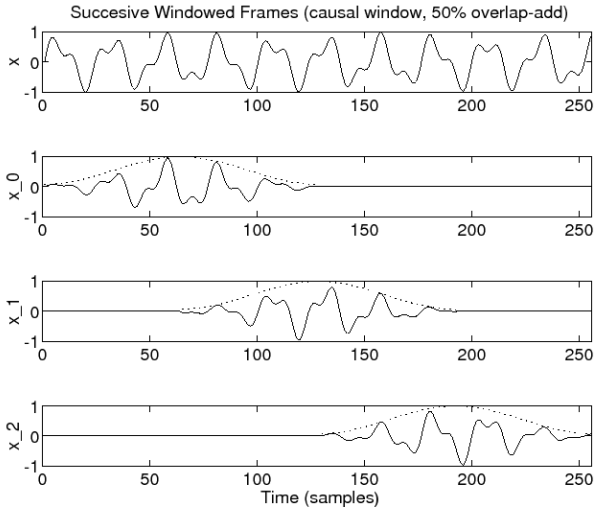
Μέρος 4. Κωδικοποίηση/Συμπίεση

Για τη συμπίεση χρειάζεται να κβαντίσουμε τις παραμέτρους του LPC συνθέτη. Δείτε και την ενότητα 8.10.3 του βιβλίου των Rabiner & Schafer, 1978, καθώς και τα εισαγωγικά για γραμμική ή λογαριθμική κβάντιση από το το αντίστοιχο Κεφ. 5. Αντί για τους συντελεστές $\{\alpha_i\}$ γραμμικής πρόβλεψης, που μπορεί να παρουσιάσουν προβλήματα αστάθειας μετά την κβάντιση τους, χρησιμοποιήστε τις παρακάτω παραμέτρους (log-area-ratios):

$$g_i = \log \frac{1-\kappa_i}{1+\kappa_i}, \quad i = 1, 2, ..., p,$$

όπου κ_i είναι οι συντελεστές ανάκλασης (PARCOR) όπως προκύπτουν από τους συντελεστές γραμμικής πρόβλεψης $\{\alpha_i\}$.
Θεωρείται ότι η συχνότητα με την οποία αναλύεται η φωνή με γραμμική πρόβλεψη είναι 100 πλαίσια το δευτερόλεπτο. Χρησιμοποιείστε:
• 5 bits για τον κβαντισμό (σε γραμμική κλίμακα) κάθε παραμέτρου log-area-ratio,
• 6 bits για τον κβαντισμό (σε γραμμική κλίμακα) της θεμελιώδους περιόδου του pitch,
• 1 bit για το αν ο ήχος είναι έμφωνος ή άφωνος, και
• 5 bits για τον κβαντισμό (σε λογαριθμική κλίμακα) του κέρδους G του μοντέλου γραμμικής πρόβλεψης.
Να υπολογίσετε τον αρχικό ρυθμό πληροφορίας (σε kbits/sec) του σήματος φωνής, αν υποθέσομε 16 bits ανά δείγμα, και τον τελικό ρυθμό πληροφορίας μετά την ανωτέρω κωδικοποίηση με κβαντισμένες παραμέτρους. Ποιος είναι ο λόγος συμπίεσης; Σχεδιάστε το φασματογράφημα του κωδικοποιημένου σήματος και δημιουργείστε ένα αρχείο “synthesis_encoded_lpc.wav". Συγκρίνετε με τα αντίστοιχα αποτελέσματα του ερωτήματος 2.2. Κβαντίστε ανάλογα τις παραμέτρους του βελτιωμένου συνθέτη φωνής που υλοποιήσατε για το ερώτημα 3.2. Ποιος είναι ο λόγος συμπίεσης; Δημιουργήστε αντίστοιχα το αρχείο “synthesis_encoded_celp.wav"

ΠΑΡΑΡΤΗΜΑ Α: Overlap-Add Ανάλυση



Σχήμα 4: Ανάλυση σε επικαλυπτόμενα πλαίσια.

Θεωρήστε την ανάλυση ενός σήματος x σε πλαίσια χρησιμοποιώντας ένα μηδενικής φάσης παράθυρο w πεπερασμένου μήκους L. Τότε μπορούμε να εκφράσουμε το m παραθυρωμένο πλαίσιο δεδομένων ως:

$$x_m[n] \triangleq x[n]w[n-mR], \quad n \in (-\infty, \infty)$$

όπου

$$R \triangleq \text{χρονικό βήμα ανάλυσης}, m \triangleq \text{αύξων δείκτης πλαισίου}$$

Το χρονικό βήμα ανάλυσης είναι ο αριθμός των δειγμάτων μεταξύ των χρόνων έναρξης διαδοχικών πλαισίων. Συγκεκριμένα, είναι ο αριθμός των δειγμάτων κατά τον οποίο μετακινούμε κάθε επόμενο παράθυρο. Στο Σχήμα 4 φαίνεται το σήμα εισόδου και τρία διαδοχικά παραθυρωμένα πλαίσια ανάλυσης χρησιμοποιώντας ένα αιτιατό παράθυρο Hamming μήκους $L = 128$ με 50% επικάλυψη ($R = L/2 = 64$).
Για να δουλέψει η ανάλυση σε πλαίσια θα πρέπει να μπορούμε να ανακατασκευάσουμε το σήμα x από τα επιμέρους επικαλυπτόμενα παράθυρα, ιδανικά με απλή πρόσθεσή τους στις αρχικές χρονικές τους θέσεις. Αυτό μπορεί να γραφτεί ως:

$$x[n] = \sum_{m=-\infty}^{\infty} x_m[n] = x[n] \sum_{m=-\infty}^{\infty} w[n - mR]$$

Οπότε, $x = \sum_m x_m$ αν και μόνο αν

$$\sum_{m \in \mathbf{Z}} w[n - mR] = 1, \forall n \in \mathbf{Z}.$$

Η συνθήκη αυτή θα πρέπει να ελέγχεται πριν την ανάλυση σε παραθυρωμένα επικαλυπτόμενα πλαίσια αν είναι αναγκαία η ανασύνθεση του συνολικού σήματος. Ελέγξτε για παράδειγμα με το script ola.m που σας δίνεται, ότι στην περίπτωση του παραθύρου Hamming, όπως ορίζεται στο MATLAB, υπάρχει πρόβλημα για $R = (L - 1)/2$ και L περιττό.

ΠΑΡΑΡΤΗΜΑ Β: Πληροφορίες σχετικά με το pitch

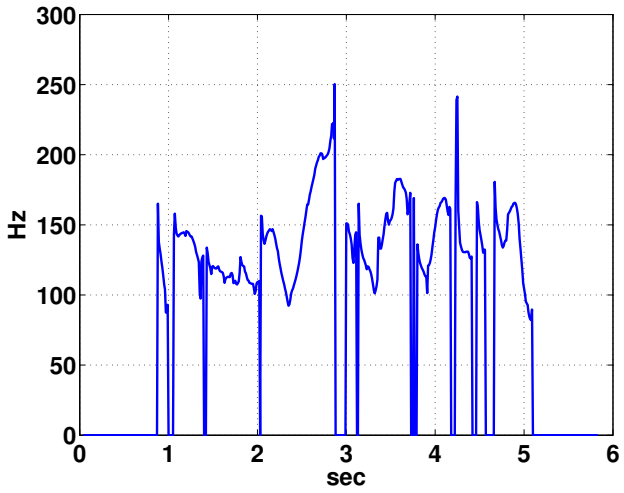
Για τη σύνθεση έμφωνων ήχων με τη χρήση του LPC μοντέλου πρέπει να εφαρμοστεί στην είσοδο κατάλληλη διέγερση η οποία στην απλούστερη περίπτωση μπορεί να θεωρηθεί ως περιοδική παλμοσειρά κρουστικών. Η περίοδος αυτής της σειράς αντιστοιχεί στη θεμελιώδη περίοδο του pitch. Για τις ανάγκες της άσκησης, σας δίνεται η θεμελιώδης συχνότητα του pitch του σήματος φωνής που σας ζητείται να ανασυνθέσετε με LPC. Οι τιμές της συχνότητας του pitch είναι σε Hz και είναι ανά 10ms, Σχήμα 5. Υπολογίστηκαν με χρήση του προγράμματος praat και εφαρμογή κατάλληλου ρυθμού δειγματοληψίας.
Επιλέγοντας ως συχνότητα της ανάλυσής σας τα 100Hz (πλαίσιο ανάλυσης διάρκειας 30ms και χρονικό βήμα ίσο με 10ms) έχετε μία τιμή της συχνότητας του pitch ανά πλαίσιο, την οποία και μπορείτε να θεωρήσετε σταθερή για όλη τη διάρκεια του πλαισίου (pitch-ασύγχρονη σύνθεση). Κατά σύμβαση, θεωρήστε ότι μηδενική τιμή συχνότητας του pitch αντιστοιχεί σε άφωνο ήχο, οπότε και θα πρέπει να επιλέξετε άλλη μορφή διέγερσης, Σχήμα 2.

ΠΑΡΑΡΤΗΜΑ Γ: Σύνθεση και Παραγωγή Φωνής, Επιπρόσθετες Σχετικές Πηγές

Διαφορετικές διεγέρσεις (αντί για απλή περιοδική σειρά κρουστικών) για σύνθεση φωνής (Rosenberg, Fant):
`http://www.asel.udel.edu/speech/tutorials/production/source.htm`

Διαφορετικοί τρόποι σύνθεσης φωνής:
`http://www.ims.uni-stuttgart.de/~moehler/synthspeech/`

Klatt Formant Synthesizer (ενδεικτικό του πόσο πολύπλοκος τελικά μπορεί να είναι ο καθορισμός των παραμέτρων για σύνθεση φωνής με χρήση του μοντέλου πηγής - φίλτρου και προκαθορισμό των formants):
`http://www.asel.udel.edu/speech/tutorials/synthesis/Klatt.html`



Σχήμα 5: Συχνότητα του pitch ανά πλαίσιο ανάλυσης του σήματος στο speech.wav

Σύνθεση φωνής με χρήση μοντέλου που περιγράφει τη γεωμετρία της φωνητικής οδού (Articulatory Synthesis)
`http://www.haskins.yale.edu/facilities/asy-demo.html`

Παραγωγή Φωνής (Βίντεο μαγνητικής τομογραφίας κατά την ομιλία):
`http://sail.usc.edu/span/index.php`