ΕΡΓΑΣΤΗΡΙΑΚΗ ΑΣΚΗΣΗ 12

Ψηφιακή Επεξεργασία Σημάτων Ομιλίας

1. Εισαγωγικά Στοιχεία

Η επεξεργασία σημάτων ομιλίας είναι μία από τις πιο γόνιμες περιοχές εφαρμογών τεχνικών της ψηφιακής επεξεργασίας σήματος. Βασικά προβλήματα επεξεργασίας ομιλίας όπως η κατάτμηση, η προέμφαση του σήματος ομιλίας, η ψηφιακή κωδικοποίηση, η αναγνώριση ομιλίας και ομιλητή αποτελούν ενδιαφέρουσες περιπτώσεις στις οποίες μπορούν να εφαρμοστούν αλγόριθμοι ψηφιακής επεξεργασίας σήματος [2, 6]. Στις εφαρμογές αυτές είναι πολύ σημαντικό να κατανοήσουμε τις ιδιότητες των σημάτων ομιλίας και να αναζητήσουμε «έξυπνους» τρόπους και σωστές χρήσεις των τεχνικών της Ψηφιακής Επεξεργασίας Σημάτων. Τα προβλήματα που παρουσιάζονται στα πλαίσια της εργασίας αυτής αποτελούν μία εισαγωγή στην επεξεργασία αυτών των σημάτων και έχουν σκοπό να σκιαγραφήσουν μερικές από τις ιδιότητές τους και να αναδείξουν την ανάγκη εφαρμογής βασικών τεχνικών της επεξεργασίας σημάτων.

Τα σήματα ομιλίας είναι ιδιαίτερα πολύπλοκα και σύνθετα, αν σκεφτούμε ότι περιέχουν τόσο την «ουσιαστική πληροφορία» που συνδέεται με κάποιο συγκεκριμένο μήνυμα όσο και πληροφορίες που σχετίζονται με τον ομιλητή όπως για παράδειγμα το φύλο, την ηλικία, την συναισθηματική κατάσταση που βρίσκεται ο ομιλών, κ.ά. Η φωνή όπως φαίνεται στο Σχήμα 1, παράγεται μέσα σε έναν αγωγό μεταβλητής διατομής που αρχίζει από τις φωνητικές χορδές και τελειώνει στα χείλη. Ένας πρόσθετος αγωγός, η ρινική κοιλότητα, συμμετέχει στην παραγωγή των ένρινων φωνημάτων. Ο

τρόπος με τον οποίο παράγονται τα σήματα ομιλίας μπορεί, σε γενικές γραμμές, να περιγραφεί ως ακολούθως:

- Η κοιλότητα του στέρνου διαστέλλεται και συστέλλεται εξωθώντας τον αέρα από τους πνεύμονες μέσω της τραχείας να περάσει την γλωττίδα (το άνοιγμα των φωνητικών χορδών).
- Από τον λαρυγγικό σωλήνα η ροή του αέρα φθάνει στην φαρυγγική κοιλότητα, κατευθυνόμενη προς την γλώσσα και φθάνει στην στοματική ή/και την ρινική κοιλότητα.
- Τέλος η ροή του αέρα απελευθερώνεται από το στόμα ή/και από την μύτη και γίνεται αντιληπτή ως ομιλία.

Η απόκριση του συστήματος παραγωγής ομιλίας είναι στενά συνδεδεμένη με την μορφή της διέγερσης που εφαρμόζεται στην είσοδό του η οποία με την σειρά της καθορίζεται από την θέση των φωνητικών χορδών. Αν και ο τρόπος με τον οποίο επηρεάζει τη μορφή της διέγερσης η σχετική θέση των φωνητικών χορδών είναι ένα θέμα το οποίο χρήζει περαιτέρω έρευνας, στα πλαίσια της εργασίας αυτής θα εστιάσουμε την προσοχή μας στις δύο περιπτώσεις που ακολουθούν [2, 6]:

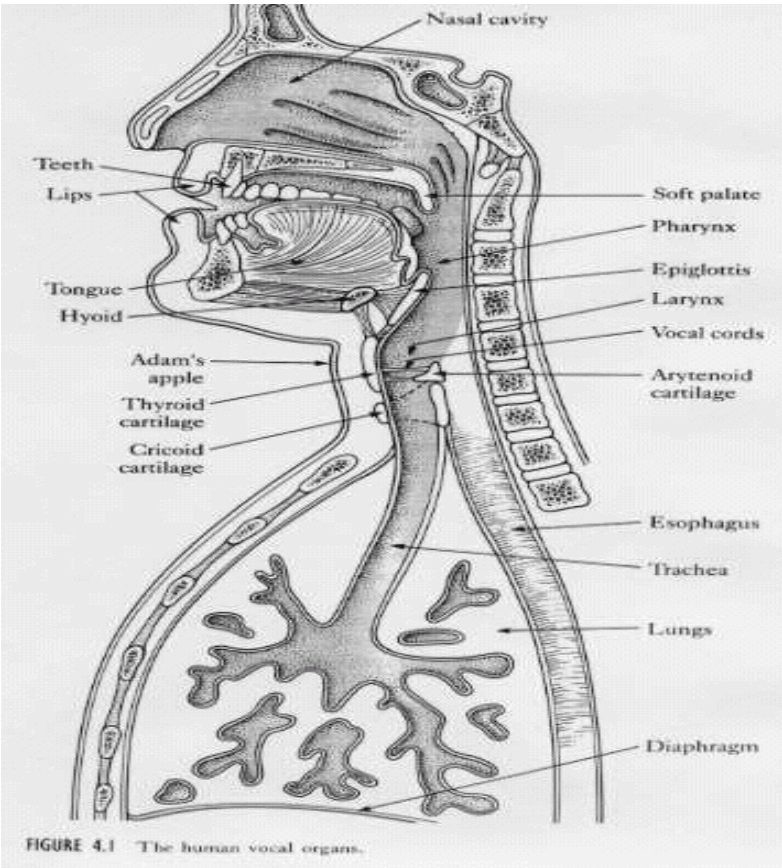
- Αν οι φωνητικές χορδές είναι τεντωμένες και σε μικρή απόσταση μεταξύ τους, η διέλευση του αέρα έχει σαν αποτέλεσμα την δημιουργία πίεσης η οποία όταν αυξηθεί σημαντικά αναγκάζει τις χορδές να αποχωριστούν. Η ροή του αέρα προς την φαρυγγική κοιλότητα, έχει σαν αποτέλεσμα την δημιουργία υποπίεσης και επομένως την επαναφορά των φωνητικών χορδών στην αρχική τους θέση. Ο κύκλος αυτός επαναλαμβάνεται με μια συγκεκριμένη συχνότητα η οποία είναι γνωστή σαν τόνος (pitch). Η παλμική αυτή διέγερση προκαλεί την ταλάντωση του φωνητικού σωλήνα στις ιδιοσυχνότητές του. Όπως θα δούμε στην συνέχεια, στο φάσμα των σημάτων ομιλίας είναι εμφανείς οι συντονισμοί του φωνητικού καναλιού (formants). Σε αυτόν τον μηχανισμό διέγερσης του φωνητικού συστήματος οφείλουν την παραγωγή τους τα φωνήματα των φωνηέντων τα οποία ονομάζονται και ηχηρά φωνήματα.

- Αν οι φωνητικές χορδές είναι απομακρυσμένες μεταξύ τους, η ροή του αέρα περνάει από την γλωττίδα ανεπηρέαστη. Αν η ροή του αέρα συναντήσει μία στένωση κατά μήκος του φωνητικού σωλήνα, δημιουργείται μία τυρβώδης ροή του αέρα που έχει τα χαρακτηριστικά θορύβου και σαν αποτέλεσμα την ακανόνιστη διέγερση του συστήματος ομιλίας. Σε αυτόν τον μηχανισμό διέγερσης του φωνητικού συστήματος οφείλουν την παραγωγή τους τα Τυρβώδη σύμφωνα (άηχα φωνήματα).

Όπως ήδη αναφέραμε υπάρχουν πολλές τεχνικές της ψηφιακής επεξεργασίας σημάτων που εφαρμόζονται στα σήματα της ομιλίας η δυσκολία των οποίων ποικίλει. Στην συνέχεια παρατίθενται μερικά βασικά προβλήματα.

2. Κατάτμηση Σήματος Ομιλίας

Η ανίχνευση φωνημάτων (οι στοιχειώδεις ήχοι που αποτελούν το σήμα ομιλίας ονομάζονται



Σχήμα 1

φωνήματα) μέσα σ' ένα σήμα ομιλίας αποτελεί ένα από τα πιο δύσκολα προβλήματα που παρουσιάζονται κατά την ψηφιακή επεξεργασία τέτοιων σημάτων. Γενικά, η διαδικασία αυτή είναι πολύ δύσκολο να πραγματοποιηθεί αυτόματα με τη βοήθεια υπολογιστή. Ακόμη και από τον ίδιο τον άνθρωπο, απαιτείται μεγάλη ικανότητα, εμπειρία και εξειδικευμένη γνώση. Παρολ' αυτά είναι πολύ εποικοδομητικό να προσπαθήσει κάποιος να αναγνωρίσει τα φωνήματα στα οποία αντιστοιχούν τα διάφορα τμήματα του σήματος. Ένα παράδειγμα ενός σήματος ομιλίας που προέκυψε κατά την προφορά της πρότασης «***Oak is strong and also gives shade***» φαίνεται στο Σχήμα 2. (Χρησιμοποιώντας κατάλληλα τις συναρτήσεις ***plot(.)*** και ***subplot(.)*** της MATLAB). Η κυματομορφή έχει δειγματοληπτηθεί με ρυθμό 8000 δείγματα / δευτερόλεπτο.

Το συγκεκριμένο σήμα ομιλίας είναι αποθηκευμένο στο αρχείο **speech_signal.mat** και μπορείτε να το φορτώσετε στην MATLAB με την εντολή ***load*** speech_signal.

2.1 Αναπαράσταση των φωνητικών συμβόλων του κειμένου

Πρώτα γράψτε την αναπαράσταση των φωνητικών συμβόλων της πρότασης «***Oak is strong and also gives shade***» με χρήση του συστήματος ARPABET που δίνεται στον Πίνακα 1.

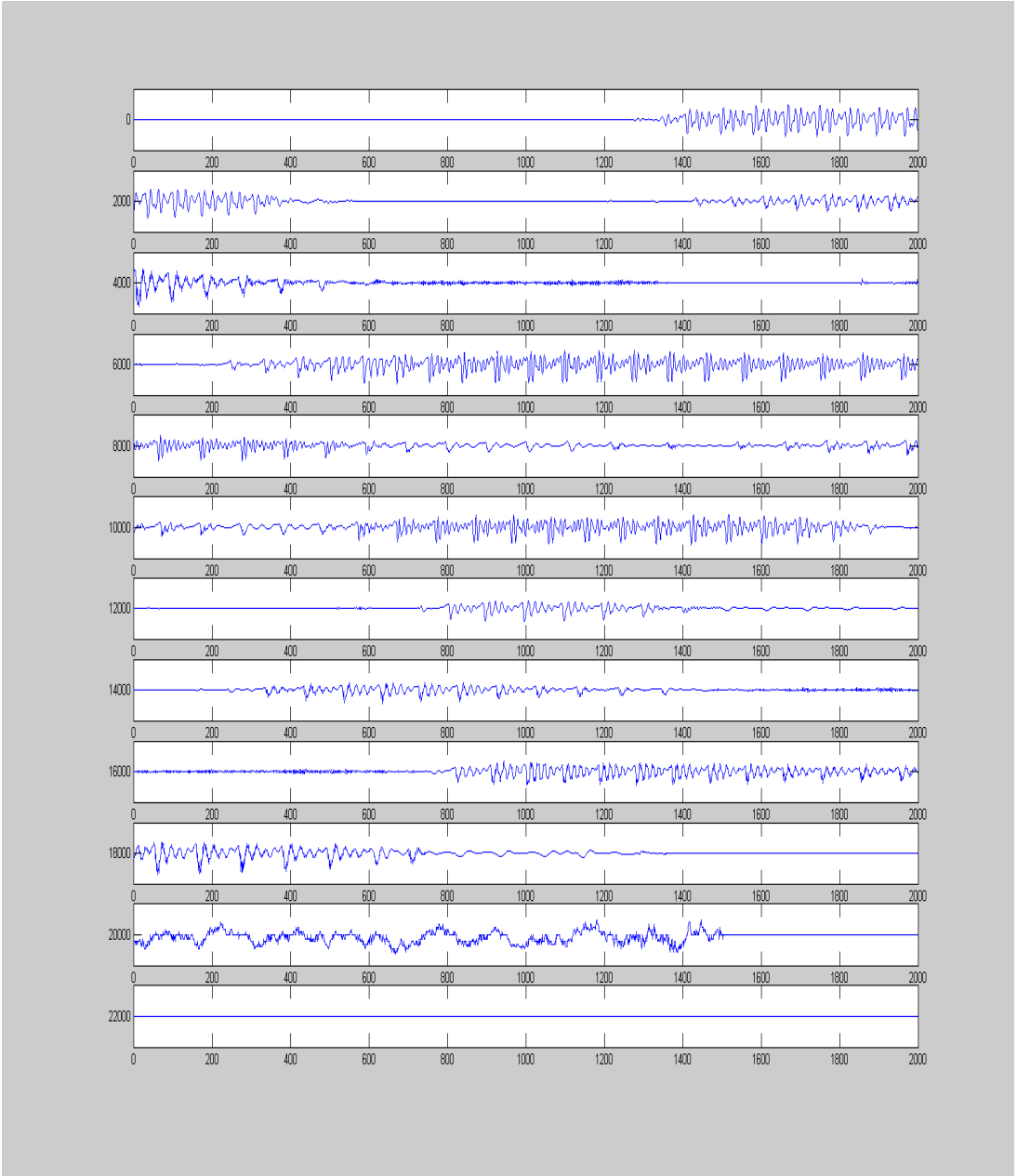
2.2 Εττικετοποίηση των φωνητικών συμβόλων (Phonetic labeling) χρησιμοποιώντας τις αναπαραστάσεις της κυματομορφής

2.2.1. Χρησιμοποιήστε την αναπαράσταση του Σχήματος 2 για να εξετάσετε το σήμα

ομιλίας που περιέχεται στο αρχείο **speech_signal.mat**. Εκτιμήστε τις θέσεις πού αρχίζουν και τελειώνουν τα φωνήματα (Απαιτείται ιδιαίτερη προσοχή για τα φωνήματα που αναγνωρίζονται ελάχιστα στις κυματομορφές. Επίσης είναι πιθανό να υπάρχουν διαστήματα «σιωπής» ή/και «θορύβου» στην αρχή ή/και στο τέλος του αρχείου).

Μη ξεχάσετε να σημαδέψετε την αρχή και το τέλος των διαστημάτων αυτών και να αποδώσετε σε κάθε διάστημα το αντίστοιχο σύμβολο ARPABET. Δημιουργήστε ένα πίνακα με τα φωνήματα μαζί με τις εκτιμήσεις σας για την θέση στην οποία αρχίζει και τελειώνει το αντίστοιχο φώνημα.

Oak is strong and also gives shade



Σχήμα 2

ARPABET	Example	ARPABET	Example	ARPABET	Example
IY	beat	AY	buy	F	fat
IH	bit	OY	boy	TH	thing
EY	bait	Y	you	S	sat
EH	bet	W	wit	SH	shut
AE	bat	R	rent	V	vat
AA	Bob	L	let	DH	that
AH	but	M	met	Z	zoo
AO	bought	N	net	ZH	azure
OW	boat	NX	sing	CH	church
UH	book	P	pet	JH	judge
UW	boot	T	ten	WH	which
AX	about	K	kit	EL	battle
IX	roses	B	bet	EM	bottom
ER	bird	D	debt	EN	button
AXR	butter	H	get	DX	batter
AW	down	HH	hat	Q	(glottal stop)

Πίνακας 1

2.2.2 Χρησιμοποιώντας τα περιεχόμενα του αρχείου **speech_signal** και του πίνακα που δημιουργήσατε στο προηγούμενο ερώτημα, δημιουργήστε δύο διανύσματα $\mathbf{v_i}$, **i**=1,2 που θα περιέχουν τα ακόλουθα:

$\mathbf{v}_1$: Oa - i - - - o - - a - - a - - o - i - - - - - a - -

$\mathbf{v}_2$: - - k - s str - ng - nd - ls - g – v e s s h – d e

Παρατηρήστε ότι το διάνυσμα $\mathbf{v}_1$ έχει μόνο τα τμήματα του σήματος ομιλίας που αντιστοιχούν στα φωνήεντα της πρότασης '***Oak is strong and also gives shade***' ενώ το διάνυσμα $\mathbf{v}_2$ μόνο τα τμήματα του σήματος ομιλίας που αντιστοιχούν στα σύμφωνα της πρότασης (αν έχουμε το διάνυσμα $\mathbf{v}_1$ και το αρχικό διάνυσμα **v** πώς μπορούμε να υπολογίσουμε εύκολα το διάνυσμα $\mathbf{v}_2$;) Χρησιμοποιώντας την εντολή ***sound***()της MATLAB και δίνοντας κατάλληλες τιμές στα ορίσματά της, ακούστε το περιεχόμενο των δύο διανυσμάτων.

Από τις αναπαραστάσεις σε μορφή κειμένου των δύο διανυσμάτων, φαίνεται να είναι ευκολότερο να αποκωδικοποιήσει κάποιος την πρόταση ***μόνο*** από τα σύμφωνα παρά ***μόνο*** από τα φωνήεντα. Ισχύει αυτή η οπτική παρατήρηση και στην ακουστική;

3. Προέμφαση του Σήματος Ομιλίας

Το φάσμα των σημάτων ομιλίας είναι εξασθενισμένο στις υψηλές συχνότητες. Αυτό οφείλεται στην επίδραση της ακτινοβολίας των χειλιών [1, 2, 6]. Γενικά είναι επιθυμητό αυτή η εξασθένιση του φάσματος στις υψηλές συχνότητες να αντισταθμίζεται. Η επεξεργασία αυτή ονομάζεται

«προέμφαση». Μία απλή και ευρέως γνωστή μέθοδος «προέμφασης» είναι το γραμμικό φιλτράρισμα του σήματος ομιλίας από ένα πρώτης τάξης φίλτρο της μορφής:

$$y[n] = x[n] - ax[n-1] \tag{1}$$

όπου $x[n]$ είναι το αρχικό σήμα ομιλίας, $y[n]$ το σήμα μετά την επεξεργασία της προέμφασης και a παράμετρος.

3.1. Υπολογίστε την κρουστική απόκριση, τη συνάρτηση μεταφοράς και την απόκριση συχνότητας του συστήματος «προέμφασης». Χρησιμοποίηστε τη συνάρτηση ***freqz()*** για να απεικονίσετε την απόκριση συχνότητας του συστήματος «προέμφασης» για τιμές της παραμέτρου a 0.5, 0.8 και 0.98. Απεικονίστε και τις τρεις αποκρίσεις στην ίδια γραφική παράσταση βαθμονομώντας κατάλληλα τον άξονα των συχνοτήτων για ρυθμό δειγματοληψίας 8KHz. Πόσο θα πρέπει να επιλεχθεί το a έτσι ώστε οι υψηλές συχνότητες να ενισχυθούν;

3.2. Χρησιμοποιώντας τις συναρτήσεις ***filter()*** και ***conv()*** της MATLAB υλοποιείστε το σύστημα «*προέμφασης*» για την τιμή του $a = 0.98$. Ποιά είναι η διαφορά στις εξόδους των δύο αυτών συναρτήσεων;

3.3. Αν δεν είναι δυνατό να επεξεργαστείτε ολόκληρο το σήμα ομιλίας τι θα πρέπει να κάνετε στα άκρα των υποτμημάτων για να υλοποιήσετε το φίλτρο «προέμφασης» για ολόκληρο το σήμα;
(Υλοποιείστε τη μέθοδο της "Επικάλυψης και Άθροισης" ή της "Επικάλυψης και Διατήρησης" [4, σελ. 56-60].**)**

3.4. Χρησιμοποιήστε τις συναρτήσεις ***subplot()*** και ***plot()*** για να απεικονίσετε κάθε τμήμα του σήματος πριν και μετά την «προέμφασή» του ($a = 0.98$). Πόσο διαφέρουν μεταξύ τους τα σήματα; Ποια χαρακτηριστικά παραμένουν αμετάβλητα από την επεξεργασία της «προέμφασης»;

3.5. Δημιουργήστε ένα διάνυσμα που θα περιέχει το αρχικό σήμα φωνής ακολουθούμενο από μισό δευτερόλεπτο σιγής (f_s=8KHz) και στη συνέχεια τοποθετήστε το σήμα που προέκυψε μετά την επεξεργασία της προέμφασης. Χρησιμοποιώντας την συνάρτηση ***sound(.)*** ακούστε το σήμα και περιγράψτε την ποιοτική διαφορά που υπάρχει μεταξύ των.

4. Χρονο-Συχνοτική Ανάλυση των Σημάτων Ομιλίας

Ένα βασικό πρόβλημα του Μετασχηματισμού Fourier (όπως και των μετασχηματισμών που βασίζονται σ' αυτόν) είναι ότι δεν παρέχει καμία πληροφορία σχετικά με τον χρόνο αλλαγής του συχνοτικού περιεχομένου ενός σήματος. Ο Μετασχηματισμός Fourier Βραχέους Χρόνου (Short Time Fourier Transform) είναι ένας μετασχηματισμός που μας επιτρέπει να μελετήσουμε την χρονική μεταβολή (αν υπάρχει) του συχνοτικού περιεχομένου ενός σήματος (για την διακριτική ικανότητα του

μετασχηματισμού και την σχετική αρχή της αβεβαιότητας που την διέπει δείτε στην [7, 8]) και ορίζεται από την σχέση [1, 5]:

$$X_n(e^{j\omega}) = e^{-j\omega n}\sum_{m=-\infty}^{\infty} w[-m]x[n+m]e^{-j\omega m} = e^{-j\omega n}\widetilde{X}_n(e^{j\omega}) \qquad (2)$$

όπου $-\infty < n < \infty$ και $0 \le \omega < 2\pi$.

Αν υποθέσουμε ότι επιθυμούμε να υπολογίσουμε τις τιμές του μετασχηματισμού στις συχνότητες $\omega_k = 2\pi k/\mathrm{N}$, $k = 0,1,...,N-1$ και ότι το παράθυρο είναι τέτοιο ώστε $w[-m] = 0, m < 0$ και $m > L-1$, η Σχέση (2) μπορεί να γραφεί ως ακολούθως:

$$X_n[k] = X_n(e^{j(2\pi/N)k}) = \sum_{m=n}^{n+L-1} w[n-m]x[m]e^{-j(2\pi/N)km} \qquad (3)$$

ή ισοδύναμα

$$X_n[k] = e^{-j(2\pi/N)kn}\sum_{m=0}^{L-1} w[-m]x[n+m]e^{-j(2\pi/N)km} = e^{-j(2\pi/N)kn}\widetilde{X}_n[k] \qquad (4)$$

όπου $\widetilde{w}[m] = w[-m]$ και

$$\widetilde{X}_n[k] = \sum_{m=0}^{L-1} \widetilde{w}[m]x[n+m]e^{-j(2\pi/N)km}\ , k = 0,1\ldots,N-1\,. \qquad (5)$$

Παρατηρήστε ότι τα $X_n[k]$ και $\widetilde{X}_n[k]$ διαφέρουν μόνο στον εκθετικό συντελεστή φάσης $e^{-j(2\pi/N)kn}$ και επομένως $\left|X_n[k]\right| = \left|\widetilde{X}_n[k]\right|$. Παρατηρήστε επίσης ότι μπορούμε να υπολογίσουμε τον STFT στο σύνολο των *N* συχνοτήτων $\omega_k = 2\pi k/\mathrm{N}$, $k = 0,1,...,N-1$ με χρήση του DFT (και επομένως με συμπλήρωση μηδενικών αν απαιτείται και τον FFT) ακολουθώντας τα παρακάτω βήματα [1]:

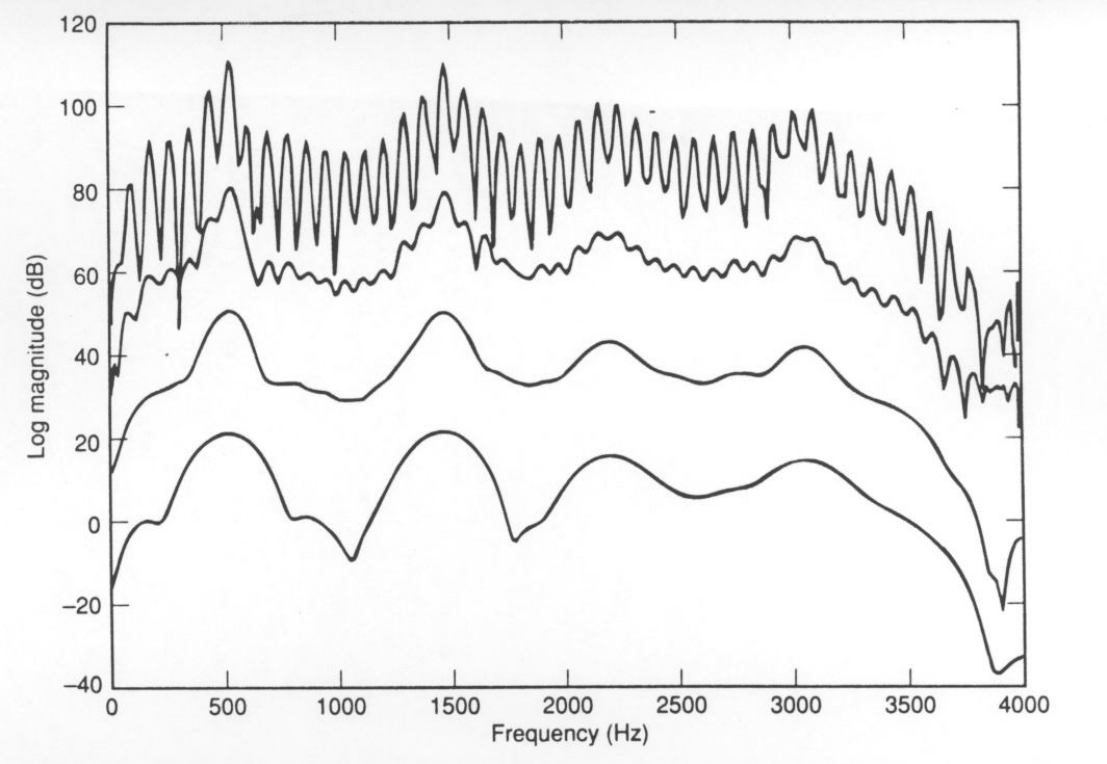
- Επέλεξε L δείγματα του σήματος αρχίζοντας από τη χρονική στιγμή n. Δηλαδή $\{x[n], x[n+1],\ldots, x[n+L-1]\}$. (Για συμμετρικά παράθυρα, είναι πιο βολικό να υποθέσουμε ότι το n είναι το κέντρο του διαστήματος του παραθύρου).
- Πολλαπλασίασε τα δείγματα του τμήματος ομιλίας με τα δείγματα του παραθύρου.
- Υπολόγισε τον DFT *N*-σημείων του πλαισίου ομιλίας μετά την «παραθύρωση» (προσθέτοντας τον αναγκαίο αριθμό μηδενικών στο τέλος του πλαισίου αν $N > L$).
- Πολλαπλασίασε με $e^{-j(2\pi/N)kn}$ (αυτό το βήμα μπορεί να παραληφθεί αν υπολογίζεται μόνο το πλάτος του STFT).
- Επανέλαβε τα τέσσερα παραπάνω βήματα για κάθε τιμή του n.

3.1. Επίδραση του Μήκους Παραθύρου

Το μήκος του παραθύρου είναι μία παράμετρος κλειδί για τον STFT. Αν το παράθυρο είναι μικρό σε σχέση με τα χαρακτηριστικά της κυματομορφής στο χρόνο, τότε ο STFT θα παρακολουθήσει τις αλλαγές αυτών των χαρακτηριστικών. Αν το παράθυρο είναι σχετικά μεγάλο, αλλαγές στο χρόνο θα αλλοιωθούν, αλλά ο STFT θα έχει καλή ανάλυση στη διάσταση της συχνότητας. Ακολουθεί ένα m-file το οποίο θα σας βοηθήσει να δείτε την επίδραση του μήκους του παραθύρου στον DFT ενός τμήματος του σήματος ομιλίας.

```
function   speccomp(x, ncenter, win, nfft, pltinc)
%    x : input signal vector
%    ncenter : sample number that windows are centered on
%    win : vector of windows to use;
%    nfft : FFT size
%     pltinc : offset of plots (in dB)
J = sqrt(-1);
x = x(:);       %--- make it a column
nwins = length(win);
X = zeros(nfft, nwins);
con = 1;
coninc = 10^(pltinc/20);
for k=1:nwins
n1 = ncenter - fix(win(k)/2);
n2 = ncenter + fix(win(k)/2);
Lh = n2-n1+1;
X(:,k) = con*fft(x(n1:n2).*hamming(Lh), nfft);
con = con/coninc;
end
f = (0:nfft/2)*(8000/nfft);
X = J*20*log10(abs(X(1:nfft/2+1,:))) + (ones(nwins,1)*f).';
plot(X)
xlabel('Frequency in Hz')
ylabel('Log Magnitude in dB')
title( 'Short-Time Spectra with Different Window Lengths')
```

Το m-file υπολογίζει τον DFT τμημάτων του σήματος ομιλίας, μετά την παραθύρωσή των. Όλα τα παραθυρώμενα τμήματα έχουν κέντρο το ίδιο δείγμα του σήματος. Όλα τα παράθυρα πρέπει να είναι περιττού μήκους για να διατηρηθεί η συμμετρία γύρω από αυτό το σημείο της κυματομορφής. Στο σχήμα που ακολουθεί φαίνεται η έξοδος αυτού του προγράμματος.



Σχήμα 3

4.1.1. Μελετήστε το παραπάνω m-file και βεβαιωθείτε ότι έχετε κατανοήσει τι κάνει.

4.1.2. Χρησιμοποιώντας διαφορετικά μήκη παραθύρου (για παράδειγμα 401,201,101,51) και $nfft = 512$ τρέξτε το m-file για το σήμα ομιλίας που περιέχεται στο αρχείο **speech_signal** επιλέγοντας ως κεντρικό σημείο τις ακόλουθες τρεις περιπτώσεις $ncenter = 3750$, 16100 και 17200.

4.1.3. Χρησιμοποιήστε τα αποτελέσματα της Ενότητας 1 για να προσδιορίσετε τα φωνήματα στα οποία αντιστοιχούν οι τρεις αυτές περιπτώσεις.

4.1.4. Ποια είναι η επίδραση της ελάττωσης του μήκους του παραθύρου;

4.1.5. Επαναλάβετε την παραπάνω διαδικασία για το σήμα που προέκυψε μετά την επεξεργασία της «προέμφασης» (Ενότητα 2) και συγκρίνετε τα αποτελέσματα με την αρχική κυματομορφή.

4.1.6. Χρησιμοποιήστε τις γραφικές παραστάσεις για να εκτιμήσετε τις τυπικές συχνότητες συντονισμού του φωνητικού καναλιού.

4.1.7. Δοκιμάστε και με άλλα τμήματα του σήματος ομιλίας και καταγράψτε τα συμπεράσματά σας.

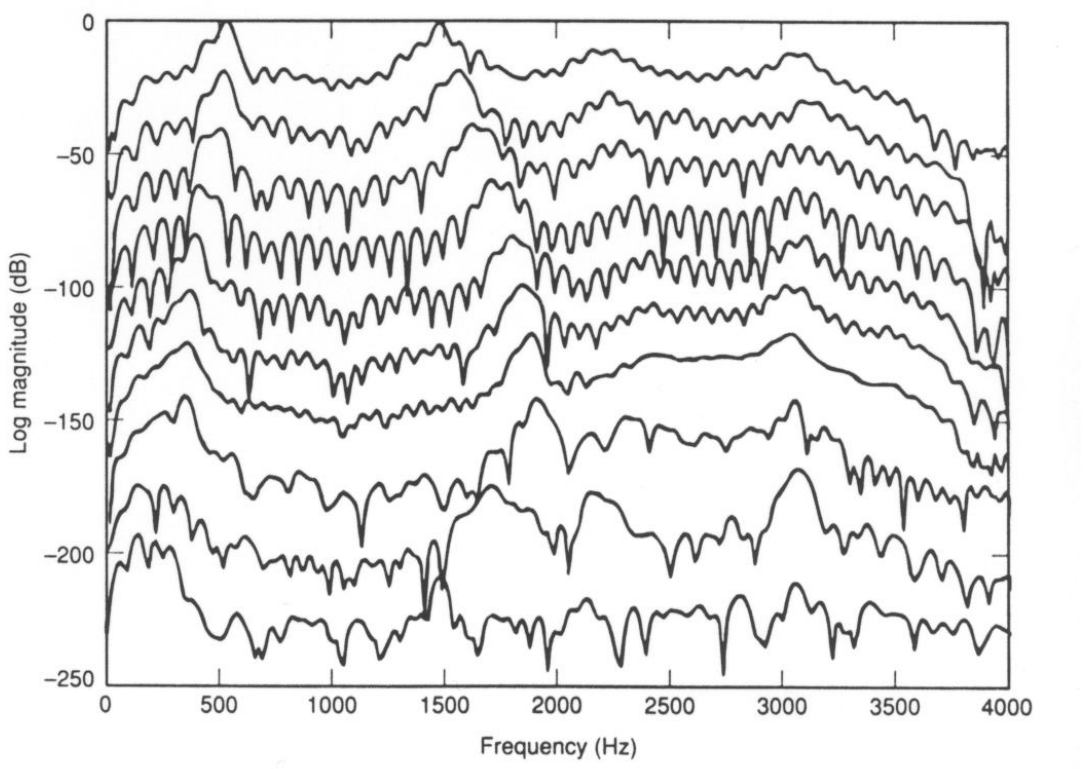
4.2. Επίδραση της θέσης του παραθύρου

Τροποποιήστε κατάλληλα το m-file *speccomp()* της προηγούμενης παραγράφου για να υπολογίσετε και να απεικονίσετε τον STFT σαν μία συνάρτηση της συχνότητας για διαφορετικές τιμές του n που ισαπέχουν. Το m-file θα πρέπει να έχει την ακόλουθη μορφή κλίσης:

```
function stspect(x, nstart, ninc, nwin, nfft, nsect, pltinc)
%       x: input signal
%       nstart: sample number that first window is centered on
%       ninc: offset between windowed segments
%       nwin: window length
%       nfft: fft size
%       nsect: number of sections to plot
%       plinc: offset of spectra in plot (in dB)
```

Το m-file θα πρέπει να δημιουργεί μία γραφική παράσταση (Σχήμα 4) όπως αυτή της προηγούμενης παραγράφου με τη συχνότητα στον οριζόντιο άξονα, άλλα αυτή τη φορά με το φάσμα να αντιστοιχεί σε διαφορετική χρονική στιγμή και όχι σε διαφορετικό μήκος παραθύρου.

1. Ελέγξτε το m-file σας για τις τρείς περιπτώσεις τις προηγούμενης παραγράφου. Χρησιμοποιήστε τιμές *nsect*=10, *ninc*=200, *nwin*=401 και *nfft*=512 και χρησιμοποιώντας τις γραφικές παραστάσεις καταγράψτε τον τρόπο με τον οποίο μεταβάλλονται με τον χρόνο οι τυπικές συχνότητες συντονισμού του καναλιού.
2. Επαναλάβετε την παραπάνω διαδικασία για το σήμα που προέκυψε μετά την επεξεργασία της «προέμφασης» (Ενότητα 2) και καταγράψτε τις διαφορές που παρατηρείτε σε σχέση με τα αποτελέσματα που πήρατε στο προηγούμενο ερώτημα.
3. Δοκιμάστε και με άλλα τμήματα του σήματος ομιλίας και καταγράψτε τα συμπεράσματά σας.



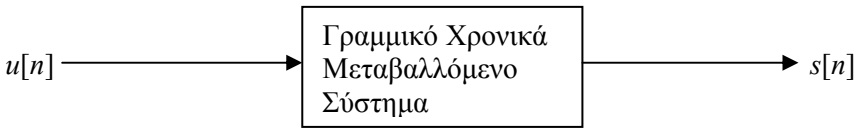
Σχήμα 4

4.3. Φασματόγραμμα

Πειραματιστείτε με την συνάρτηση ***specgram()*** *της MATLAB*. Χρησιμοποιήστέ την για να απεικονίσετε τα φασματογράμματα που προκύπτουν για τις περιπτώσεις που μελετήσατε στα προηγούμενα ερωτήματα. Σχολιάστε τα αποτελέσματά σας.

5. Ανάλυση Σημάτων Ομιλίας

Είναι φανερό ότι η παραγωγή φωνής είναι ένα σύνθετο φαινόμενο. Ιδανικά θα επιθυμούσαμε να έχουμε μοντέλα αναπαράστασης που να είναι γραμμικά και χρονικά αμετάβλητα. Δυστυχώς (ή ευτυχώς;) ο μηχανισμός παραγωγής της ανθρώπινης φωνής, όπως είδαμε στις προηγούμενες ενότητες, δεν ικανοποιεί με ακρίβεια καμία από τις παραπάνω ιδιότητες. Σε πολλές εφαρμογές της επεξεργασίας φωνής είναι θεμελιώδους σημασίας η δυνατότητα να αναπαρίσταται το σήμα ομιλίας από ένα μικρό αριθμό γραμμικών χρονικά μεταβαλλόμενων συστημάτων, όπως αυτό του σχήματος που ακολουθεί.



Σχήμα 5

όπου η είσοδος $u[n]$ είναι είτε λευκός θόρυβος (στην περίπτωση των άηχων φωνημάτων) είτε μια σχεδόν περιοδική ακολουθία παλμών (στην περίπτωση των ηχηρών φωνημάτων). Το γραμμικό σύστημα θεωρούμε πως είναι χρονικά μεταβαλλόμενο με την μεταβολή όμως να συμβαίνει αργά ώστε σε μικρά χρονικά διαστήματα (που καθορίζουν το μέγεθος του πλαισίου στα 20ms-30ms για σήματα ομιλίας που έχουν δειγματοληπτηθεί με συχνότητα 8KHz) να μπορεί να θεωρηθεί ένα χρονικά αμετάβλητο μοντέλο του πλαισίου ομιλίας που περιγράφει τα σημαντικότερα χαρακτηριστικά του σήματος. Η εξαγωγή αυτών των σημαντικότερων χαρακτηριστικών ενός σήματος ομιλίας είναι ο σκοπός των τεχνικών ανάλυσης [3, 6].

5.1 Γραμμική Πρόβλεψη και Ανάλυση Σημάτων Ομιλίας

Η ανάλυση των σημάτων ομιλίας με γραμμική πρόβλεψη είναι μια από τις πιο γνωστές τεχνικές που χρησιμοποιούνται για το σκοπό αυτό. Σύμφωνα με την τεχνική αυτή το σήμα ομιλίας $s[n]$ είναι μία διαδικασία αυτοπαλινδρόμησης τάξης p που προκύπτει από την έξοδο ενός IIR συστήματος με συνάρτηση μεταφοράς:

$$H(z)=\frac{G}{A(z)} \tag{6.1}$$

όπου

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{6.2}$$

το φίλτρο σφαλμάτων πρόβλεψης και G μια σταθερά κέρδους. Μπορεί να διαπιστωθεί εύκολα ότι η είσοδος και η έξοδος ενός τέτοιου συστήματος συνδέονται μέσω της ακόλουθης εξίσωσης διαφορών:

$$s[n] = -\sum_{k=1}^{p} a_k s[n-k] + Gu[n] \tag{7}$$

όπου $u[n]$ η διέγερση που εφαρμόζεται στην είσοδο του συστήματος.

Η τεχνική της γραμμικής πρόβλεψης αποσκοπεί στην ανεύρεση των συντελεστών a_k, $k = 1,..., p$ ώστε να ελαχιστοποιείται μία συνάρτηση κόστους η οποία βασίζεται στην ακολουθία σφαλμάτων πρόβλεψης. Γνωστές μέθοδοι οι οποίες βασίζονται στην γραμμική πρόβλεψη και οι οποίες διαφοροποιούνται ως προς τις διαφορετικές υποθέσεις που κάνουμε για την μορφή του σήματος εκτός των χρονικών ορίων του σήματος που έχουμε στην διαθέσή μας και επομένως στο πεδίο ορισμού της συνάρτησης κόστους, είναι αυτή της αυτοσυσχέτισης και αυτή της συνδιασποράς [1, 2, 3]. Η μέθοδος που θα χρησιμοποιήσουμε στα πλαίσια της εργασίας αυτής είναι αυτή της αυτοσυσχέτισης κατά την οποία ελαχιστοποιείται το τετραγωνικό σφάλμα πρόβλεψης $E = \sum_{n=1}^{N} f[n]^2$, όπου $f[n]$ το προς τα εμπρός σφάλμα εκτίμησης μεταξύ του σήματος ομιλίας $s[n]$ και του προβλεπόμενου σήματος $\hat{s}[n]$ για όλα τα διαθέσιμα δείγματα (N) του τμήματος του σήματος ομιλίας.

Στην περίπτωση αυτή η ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος ως προς τους συντελεστές του συστήματος, οδηγεί στις γνωστές εξισώσεις των Yule Walker [4, σελ. 220-222] από την λύση των οποίων προσδιορίζονται οι παράμετροι του μοντέλου αυτοπαλινδρόμησης. Η συνάρτηση ***lpc***() της MATLAB μπορεί να χρησιμοποιηθεί για την λύση του παραπάνω προβλήματος ελαχιστοποίησης και τον προσδιορισμό των βέλτιστων συντελεστών του μοντέλου της γραμμικής πρόβλεψης.

Μια βασική υπόθεση της τεχνικής της γραμμικής πρόβλεψης είναι πως οι συντελεστές του προγνώστη ταυτίζονται με εκείνους του μοντέλου ομιλίας. Συνεπώς, από τον ορισμό του μοντέλου, η έξοδος του γραμμικού φίλτρου πρόβλεψης θα δίνεται από την ακόλουθη σχέση:

$$f[n] = s[n] - \hat{s}[n] = s[n] + \sum_{k=1}^{p} a_k s[n-k] = Gu[n] \tag{8}$$

Η παραπάνω σχέση μπορεί να χρησιμοποιηθεί για τον προσδιορισμό της παραμέτρου κέρδους G. Πράγματι αν E_{min} η ελάχιστη τιμή του τετραγωνικού σφάλματος πρόβλεψης, τότε η τιμή της σταθεράς κέρδους μπορεί να οριστεί από την ακόλουθη σχέση:

$$G = \sqrt{\frac{E_{\min}}{\sum_{n=1}^{N} u[n]^2}} \tag{9}$$

5.2 Ανίχνευση και Εκτίμηση της Θεμελιώδους Συχνότητας (Τόνου)

Όπως έχουμε ήδη αναφέρει, τα ηχηρά τμήματα κάθε σήματος ομιλίας προκύπτουν από σχεδόν περιοδικές διεγέρσεις των οποίων η περίοδος καθορίζει την θεμελιώδη συχνότητα ή τόνο. Η ανίχνευση και η εκτίμηση της θεμελιώδους συχνότητας αποτελεί ένα πολύ δύσκολο πρόβλημα. Συγκεκριμένα, υποθέτοντας ότι έχουμε υπολογίσει τις βέλτιστες παραμέτρους του μοντέλου γραμμικής πρόβλεψης και την σταθερά κέρδους αυτό που θέλουμε να κάνουμε στην συνέχεια είναι να αποφασίσουμε αν το συγκεκριμένο τμήμα της ομιλίας αντιστοιχεί σε ηχηρό ή άηχο φώνημα. Για το σκοπό αυτό χρησιμοποιώντας την ακολουθία των σφαλμάτων πρόβλεψης $f[n]$ υπολογίζουμε τον αμερόληπτο εκτιμητή[1] της ακολουθίας αυτοσυσχέτισης από την σχέση:

$$\hat{r}_{ff}[k] = \frac{1}{N-k}\sum_{n=1}^{N-k} f[n]f[n+k],\ k = 0,1,2,...,N-1 \qquad (10)$$

και συγκρίνουμε την κορυφή της κανονικοποιημένης ακολουθίας αυτοσυσχέτισης $\frac{\hat{r}_{ff}[k]}{\hat{r}_{ff}[0]},\quad k = 1,2,...,N-1$ με μία τιμή κατωφλίου T (συνήθως η τιμή του κατωφλίου είναι η 0.5).

Αν η τιμή της κορυφής είναι μικρότερη από την τιμή κατωφλίου το τμήμα θεωρείται ως άηχο και επομένως η διέγερση του συστήματος θα είναι μια διαδικασία λευκού θορύβου διασποράς 1. Διαφορετικά θεωρείται ηχηρό και η εκτίμηση της θεμελιώδους συχνότητας είναι ίση με την θέση της κορυφής.

5.2.1. Για κάθε πλαίσιο ενός φωνήματος εκτελέστε την ακόλουθη διαδικασία:

1. Χρησιμοποιήστε την συνάρτηση ***lpc()*** για να υπολογίσετε τους συντελεστές πρόβλεψης ενός προγνώστη δωδέκατης τάξης (p=12). Χρησιμοποιήστε το παράθυρο Hamming [4, σελ. 33] για την παραθύρωση του πλαισίου.
2. Σχεδιάστε σε κοινή γραφική παράσταση τα μέτρα των αποκρίσεων συχνότητας του φίλτρου πρόβλεψης σφάλματος και του μοντέλου της φωνής. Επιπροσθέτως, με την βοήθεια της συνάρτησης ***zplane***() τα μηδενικά του φίλτρου σφαλμάτων πρόβλεψης (Σχέση (6.2)). Τι παρατηρείτε αναφορικά με την σχέση μεταξύ των μηδενικών του φίλτρου αυτού και καθενός από τα παρακάτω:
    - των πόλων του μοντέλου της ομιλίας
    - των συχνοτήτων που εμφανίζονται τα μέγιστα της απόκρισης συχνότητας του μοντέλου της ομιλίας και
    - των συχνοτήτων που εμφανίζονται τα ελάχιστα της απόκρισης συχνότητας του φίλτρου σφαλμάτων πρόβλεψης.

[1] Υποθέστε ότι επιθυμούμε να εκτιμήσουμε την τιμή μιας παραμέτρου θ χρησιμοποιώντας έναν εκτιμητή $\hat{\theta}$ που είναι μια συνάρτηση των παρατηρήσεων μας. Θα λέμε ότι ο εκτιμητής μας είναι αμερόπτητος αν $E\{\hat{\theta}\} = \theta$ όπου $E\{.\}$ ο τελεστής αναμενόμενης τιμής. Με την έννοια αυτή οι εκτιμητές $\hat{r}_{ff}[k],\quad k = 0,1,2,...,N-1$ της προς εκτίμηση ακολουθίας αυτοσυσχέτισης $r_{ff}[k],\quad k = 0,1,2,...,N-1$ μπορεί εύκολα να αποδειχθεί ότι είναι αμερόληπτοι.

3. Υπολογίστε την έξοδο του φίλτρου πρόβλεψης και την ακολουθία σφαλμάτων πρόβλεψης $f[n]$. Σχεδιάστε σε κοινή γραφική παράσταση:
    - το αρχικό πλαίσιο του σήματος ομιλίας
    - το παραθυρωμένο πλαίσιο και
    - την ακολουθία σφαλμάτων πρόβλεψης

    σχολιάστε τα αποτελέσματά σας.
4. Χρησιμοποιώντας τις Σχέσεις (9) και (10), εκτιμήστε τις τιμές της σταθεράς κέρδους *G* και της Θεμελιώδους Συχνότητας του πλαισίου.

5.2.2.1. Το φώνημα «sh» της λέξης «shade» αρχίζει κοντά στο δείγμα 15500 και τελειώνει περίπου στο δείγμα 16750. Για κάθε πλαίσιο του φωνήματος εκτελέστε τα Βήματα 1-4 και καταγράψτε τα αποτελέσματά σας.

5.2.2.2. Να επαναλάβετε το προηγούμενο ερώτημα με το σήμα ομιλίας όμως να έχει υποστεί την επεξεργασία της «προέμφασης». Σχολιάστε τα αποτελέσματα σας.

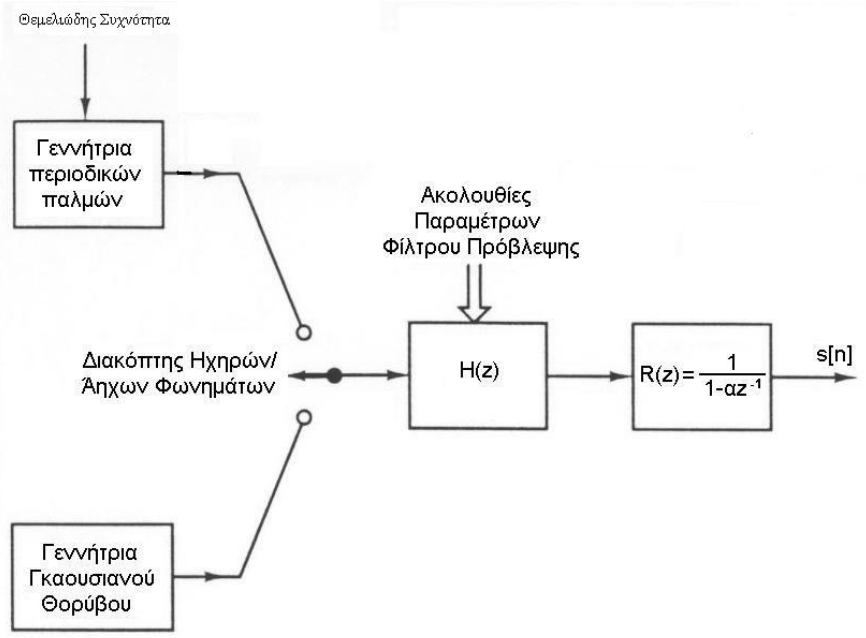
5.2.2.3. Επαναλάβετε την παραπάνω διαδικασία για το φώνημα «aa» της ίδιας λέξης το οποίο αρχίζει κοντά στο δείγμα 16750 και τελειώνει κοντά στο δείγμα 18800.

5.2.2.4. Τι παρατηρείτε σχετικά με τις διαφορές μεταξύ των φωνημάτων «aa» και «sh»;

5.2.2.5. Επαναλάβετε την παραπάνω διαδικασία χρησιμοποιώντας διαφορετικές τάξεις του προγνώστη. Για παράδειγμα σχεδιάστε την απόκριση συχνότητας ενός προγνώστη μήκους *p* ίσο με 8, 10, 12 και 24. Για κάθε προγνώστη υπολογίστε το τετραγωνικό σφάλμα πρόβλεψης και σχολιάστε τα αποτελέσματά σας.

6. Σύστημα Σύνθεσης Ομιλίας (Προαιρετικό)

Χρησιμοποιώντας τα αποτελέσματα της προηγούμενης παραγράφου και το σύστημα του Σχήματος 6, συνθέστε τουλάχιστον μία λέξη της πρότασης που περιέχεται στο αρχείο **speech_signal**.



Σχήμα 6

Βιβλιογραφία

[1] C.S. Burrus, J.H. McClellan, A.V. Oppenheim, T.W. Parks, R.W. Schafer, and H.W. Schussler "Computer-Based Exercises for Signal Processing Using MATLAB," Prentice-Hall, 1994.

[2] B. Gold and N. Morgan, “Speech and Audio Signal Processing,” John Wiley & Sons, Inc., 2000.

[3] M. H. Hayes, “Statistical Digital Signal Processing and Modeling,” John Wiley & Sons, Inc., 1996.

[4] Γ. Β. Μουστακίδης, "Βασικές Τεχνικές Ψηφιακής Επεξεργασίας Σημάτων," Εκδόσεις Τζιόλα , 2004.

[5] Sanjit K. Mitra, "Digital Signal Processing, A Computer-Based Approach," McGraw-Hill, 1998.

[6] P. E. Papamichalis, “Practical Approaches to Speech Coding,” Prentice-Hall 1987.

[7] A. Papoulis, “Signal Analysis,” McGraw-Hill, 1985.

[8] M. Veterli and J. Kovacevic, “Wavelets and Subband Coding,” Prentice Hall Inc. 1995.

[9] Matlab's Tutorial, Mathworks Inc.