ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

ΠΑΡΑΡΤΗΜΑ ΝΑΥΠΑΚΤΟΥ

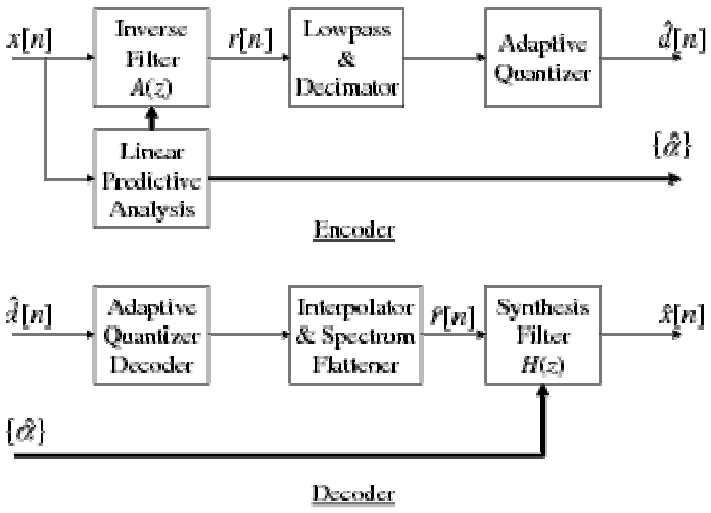
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ Τ.Ε

(Πρώην ΤΕΣΥΔ)

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΜΕ ΘΕΜΑ:

ΓΡΑΜΜΙΚΗ ΠΡΟΒΛΕΠΤΙΚΗ ΚΩΔΙΚΟΠΟΙΗΣΗ ΦΩΝΗΣ

(LINEAR PREDICTIVE CODING OF SPEECH)



Σπουδαστής: Παπάς Σπύρος Επιβλέπων: Παρασκευάς Μιχαήλ

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 2

ΚΕΦΑΛΑΙΟ 3

ΠΡΟΛΟΓΟΣ

Ευχαριστώ θερμά την οικογένεια μου για όλη την υποστήριξη τόσο κατά την διάρκεια των σπουδών μου όσο και κατά την εκπόνηση αυτής της εργασίας. Ευχαριστώ τον επιβλέποντα καθηγητή κ. Παρασκευά Μιχαήλ για την πολύτιμη βοήθεια που μου έδωσε και την καθοδήγηση του κατά την διάρκεια της προετοιμασίας αυτής της εργασίας. Επίσης, ευχαριστώ όλους τους καθηγητές του τμήματος Τηλεπικοινωνιακών Συστημάτων Και Δικτύων για τις γνώσεις που μου έδωσαν.

Τέλος θα ήθελα να ευχαριστήσω τους φίλους μου και τους δικούς μου ανθρώπους που πήραν μέρος σε αυτή την εργασία με κάθε τρόπο. Χωρίς την βοήθειά τους το αποτέλεσμα δεν θα ήταν το ίδιο. .

ΕΙΣΑΓΩΓΗ

Υπάρχουν πολλοί διαφορετικοί τύποι συμπίεσης ομιλίας (speech compression) που κάνουν χρήση μιας ποικιλίας διαφορετικών τεχνικών. Ωστόσο, οι περισσότερες μέθοδοι συμπίεσης ομιλίας εκμεταλλεύονται το γεγονός ότι η παραγωγή ομιλίας συμβαίνει μέσω αργών ανατομικών κινήσεων και ότι η παραγόμενη ομιλία έχει ένα περιορισμένο εύρος συχνότητας. Η συχνότητα της παραγόμενης ανθρώπινης ομιλίας κυμαίνεται από περίπου 300 Hz έως 3400 Hz. Η συμπίεση ομιλίας αναφέρεται συχνά ως κωδικοποίηση ομιλίας (speech coding) η οποία ορίζεται ως μια μέθοδος για τη μείωση της ποσότητας των πληροφοριών που απαιτούνται για την αναπαράσταση ενός σήματος ομιλίας. Οι περισσότερες μορφές κωδικοποίησης ομιλίας βασίζονται συνήθως σε έναν αλγόριθμο με απώλειες (lossy algorithm). Οι αλγόριθμοι με απώλειες θεωρούνται αποδεκτοί όταν κωδικοποιούν την ομιλία διότι η απώλεια ποιότητας είναι συχνά μη ανιχνεύσιμη από το ανθρώπινο αυτί.

Υπάρχουν πολλά άλλα χαρακτηριστικά για την παραγωγή λόγου που μπορούν να αξιοποιηθούν από τους αλγόριθμους κωδικοποίησης ομιλίας. Ένα γεγονός που χρησιμοποιείται συχνά είναι ότι η περίοδος σιωπής καταλαμβάνει περισσότερο από 50% μιας συνομιλίας. Ένας εύκολος τρόπος να μειωθεί το εύρος ζώνης και το ποσό των πληροφοριών που χρειάζονται για να αναπαρασταθεί ένα σήμα ομιλίας είναι να μην μεταδίδεται η σιωπή. Ένα άλλο γεγονός σχετικά με την παραγωγή του λόγου που μπορεί να ληφθεί ως πλεονέκτημα είναι ότι μηχανικά υπάρχει υψηλή συσχέτιση μεταξύ γειτονικών δειγμάτων του λόγου. Οι περισσότερες μορφές της συμπίεσης ομιλίας επιτυγχάνονται με τη μοντελοποίηση της διαδικασίας της παραγόμενης ομιλίας ως γραμμικό ψηφιακό φίλτρο. Το ψηφιακό φίλτρο και οι παράμετροί του συνήθως κωδικοποιούνται ώστε να επιτευχθεί συμπίεση του σήματος ομιλίας.

Η Γραμμική Προβλεπτική Κωδικοποίηση Φωνής (Linear Predictive Coding ή LPC) είναι μια από τις μεθόδους συμπίεσης που μοντελοποιεί τη διαδικασία της παραγωγής ομιλίας.

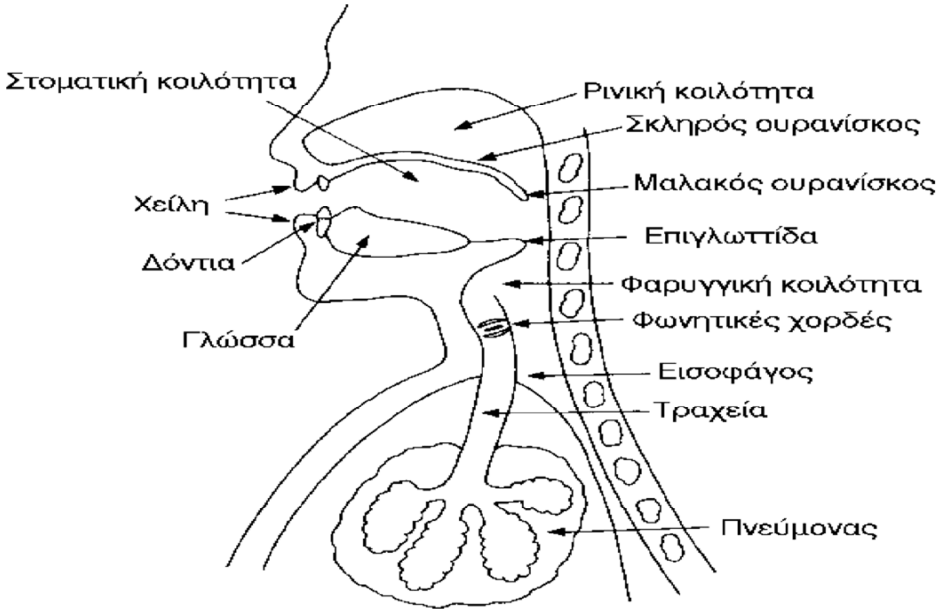
ΚΕΦΑΛΑΙΟ 1

ΤΟ ΣΗΜΑ ΟΜΙΛΙΑΣ : ΠΑΡΑΓΩΓΗ, ΑΝΤΙΛΗΨΗ ΚΑΙ ΤΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΤΟΥ

Για να μπορέσει να αναπτυχθεί ένας αλγόριθμος αναγνώρισης φωνής είναι σημαντικό να μελετηθεί πρώτα ο τρόπος παραγωγής της φωνής, καθώς επίσης και το ακουστικό σύστημα του ανθρώπου, μιας και πρέπει να γνωρίζουμε καλά την μορφή και τα χαρακτηριστικά του σήματος που θα πρέπει να επεξεργαστούμε. Τα διάφορα στοιχεία της ανθρώπινης φυσιολογίας που θα δούμε εδώ, βοηθούν στο να κατανοηθεί ο τρόπος που ο εγκέφαλος καταλαβαίνει και ερμηνεύει τα ακουστικά ερεθίσματα, ούτος ώστε να μπορέσουμε να εξομοιώσουμε τις λειτουργίες αυτές με τεχνητά μέσα και να φτάσουμε σε όσο το δυνατόν καλύτερο αποτέλεσμα.

Η ανθρώπινη ομιλία

Οι ήχοι της ομιλίας δημιουργούνται από τον αέρα που εκπνέουν οι πνεύμονες, παράγοντας είτε ταλαντώσεις των φωνητικών χορδών για τα φωνήεντα είτε στροβιλισμούς του αέρα σε κάποιο σημείο της φωνητικής περιοχής (πχ. Δόντια) για τα σύμφωνα. Οι ήχοι που παράγονται επηρεάζονται από τα σχήματα που παίρνει η φωνητική κοιλότητα με αποτέλεσμα να δημιουργούνται διάφορες αρμονικές του αρχικού σήματος. Τα διάφορα όργανα που συμμετέχουν στην παραγωγή των ήχων της φωνής φαίνονται στην εικόνα 1.1.

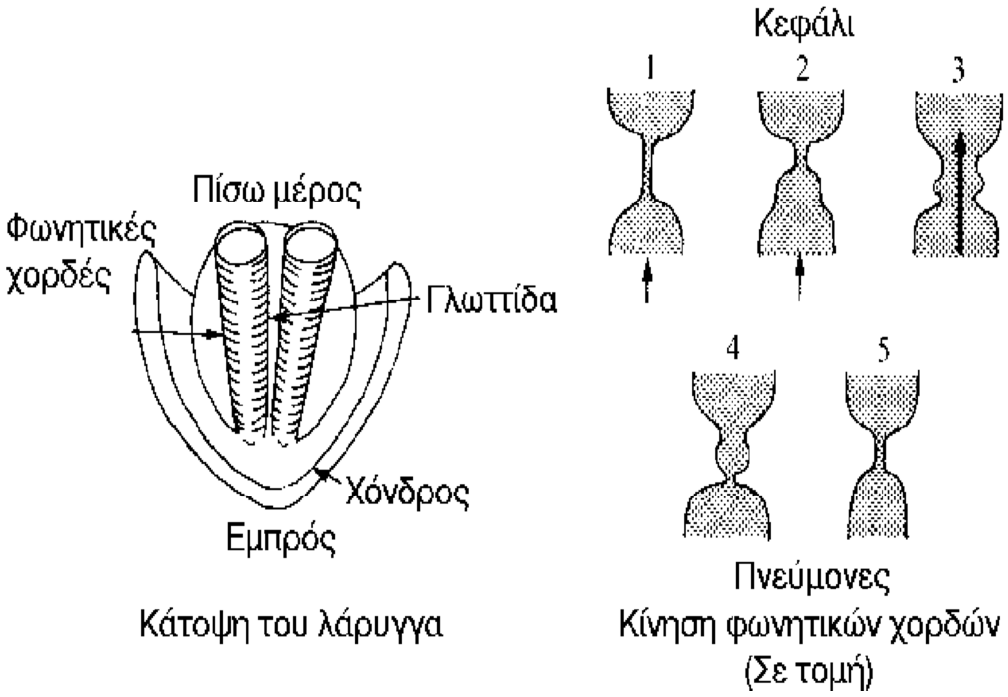


Εικόνα 1.1 Τα όργανα ομιλίας του ανθρώπου

Τα ανθρώπινα όργανα ομιλίας

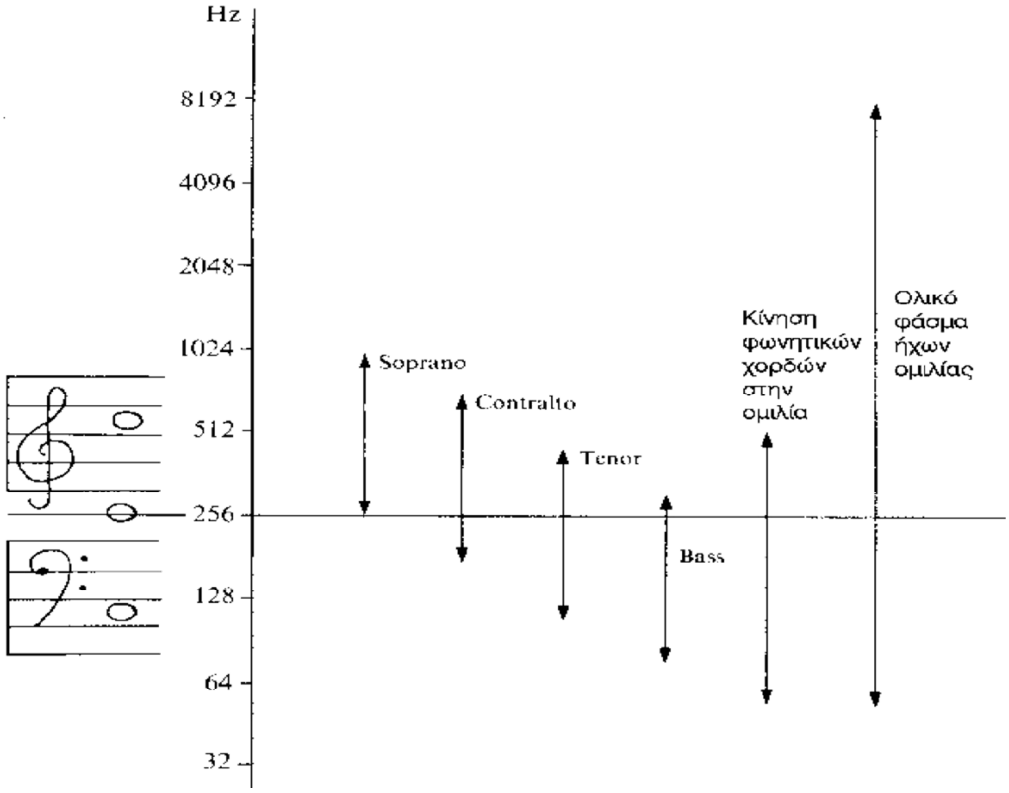
Τα κυριότερα σημεία του ανθρώπινου συστήματος ομιλίας είναι οι πνεύμονες, ο λάρυγγας, οι φωνητικές χορδές, η γλώσσα και τα δόντια. Δυο είναι οι βασικές μέθοδοι παραγωγής ήχων. Στην πρώτη οι φωνητικές χορδές που βρίσκονται στο λάρυγγα ταλαντώνονται σε σταθερές συχνότητες από τον αέρα που εκπνέουν οι πνεύμονες, οπότε παράγονται τα ηχηρά φωνήεντα. Η δεύτερη μέθοδος παράγει ήχους από τους στροβιλισμούς του αέρα σε κάποιες από τις φωνητικές περιοχές όπως τα δόντια ή τα χείλη και παράγει τα σύμφωνα.

Οι φωνητικές χορδές είναι ελαστικές, οπότε ανοίγουν μετά από κάθε παλμό της γλωττίδας λόγω της πίεσης του αέρα, και μαζεύουν όταν η πίεση χαμηλώσει (λόγω του φαινομένου Bernoulli). Αν δούμε τις χορδές σε κάθετη τομή όπως στην εικόνα 1.2 θα δούμε ότι οι φωνητικές χορδές δεν ανοιγοκλείνουν ομοιόμορφα, αλλά κυματίζουν από κάτω προς τα πάνω.



Εικόνα 1.2 Οι κινήσεις των φωνητικών χορδών κατά την ομιλία

Η συχνότητα ταλάντωσης εξαρτάται από την τάση που ασκούν οι μύες, τη μάζα και το μήκος των φωνητικών χορδών. Το μήκος των χορδών στους άνδρες είναι από 17 ως 24 mm, ενώ στις γυναίκες από 13 ως 17 mm. Η μέση βασική συχνότητα των παλμών της γλωττίδας είναι για τους άνδρες περίπου 125Hz, για τις γυναίκες 200Hz και για τα παιδιά 300 Hz. Η εικόνα 1.3 δείχνει τις βασικές συχνότητες που παράγονται από διάφορες φωνές τραγουδιστών σε σχέση με τις νότες και την ομιλία.



Εικόνα 1.3 Οι συχνότητες που παράγονται κατά την ομιλία

Οι φωνητικές χορδές παράγουν και αρμονικές πολλαπλάσιες της βασικής αρμονικής. Το πλάτος των αρμονικών μειώνεται όσο αυξάνονται οι συχνότητες.

Κάθε αλλαγή στους μύες του προσώπου του ομιλητή, επηρεάζει την βασική αρμονική που παράγεται από τις φωνητικές χορδές. Η φωνητική κοιλότητα έχει έντονη κινητικότητα και αλλάζει γεωμετρία ανάλογα με την θέση της γλώσσας, του φάρυγγα, των χειλιών και των μαγούλων. Η αναπνευστική οδός είναι πιο σταθερή αλλά μπορεί να συνδέεται ακουστικά με την φωνητική περιοχή ανάλογα με την θέση που τοποθετείται ο ουρανίσκος.

Φθόγγοι

Το μικρότερο στοιχείο της φωνής είναι οι φθόγγοι. Οι φθόγγοι συμβολίζονται διεθνώς (κατά IPA, International Phonetic Alphabet) με μερικά γράμματα ή σύμβολα που γράφονται ανάμεσα σε δύο κάθετες γραμμές όπως /p/ για τη λέξη "pan". Στην πραγματικότητα η προφορά του φθόγγου /p/ δεν είναι η ίδια πάντα αλλά αλλάζει ανάλογα με τη θέση του μέσα στη λέξη όπως στις αγγλικές λέξεις "pan" και "span". Στην αγγλική γλώσσα υπάρχουν 40 φθόγγοι που αρκούν για να περιγράψουν όλους τους φωνητικούς ήχους της ομιλίας. Ο πίνακας των διαφόρων φθόγγων φαίνεται στον πίνακα 1.1.

Our symbol	IPA symbol	Key word	Our symbol	IPA symbol	Key word
Vowels			*Fricatives*		
/ee/	i	each	/f/	f	free
/i/	ɪ	it	/th/	θ	thin
/e/	ɛ	end	/s/	s	see
/ar/	ą	hard	/sh/	ʃ	shall
/u/	ʊ	good	/v/	v	vine
/uu/	u	ooze	/dh/	ð	then
/er/	ɜ	bird (neutral)	/z/	z	zoo
/aa/	æ	had	/xh/	ʒ	azure
/a/	ʌ	bud	/h/	h	he
/aw/	ɔ	hoard	*Affricates*		
/(ə)/	ə	allow (schwa)	/tsh/	tʃ	chair
/Q/	ɒ	hot	/dzh/	dʒ	jar
Diphthongs			*Semi vowels*		
/ei/	ei	aid	/w/	w	we
/u(ə)/	ʊə	pure	/y/	j	you
/au/	au	cow	/r/	r	red
/ou/	əu	own	/l/	l	live
Plosives			*Nasals*		
/p/	p	pie	/m/	m	me
/t/	t	ten	/n/	n	no
/k/	k	key	/ng/	ŋ	sing
/b/	b	be			
/d/	d	den			
/g/	g	go			

Πίνακας 1.1 Η γραφή των φθόγγων της Αγγλικής γλώσσας με το διεθνές φωνητικό αλφάβητο

Τα φωνήεντα /a/, /e/, / i/ παράγονται από τις ταλαντώσεις των φωνητικών χορδών που βρίσκονται στην κορυφή της τραχείας. Η βασική τους συχνότητα εξαρτάται και από την ροη του αέρα, αλλά κυρίως από την δύναμη που εξασκείται στις χορδές. Το /@/ δεν είναι κάποιο συγκεκριμένο φωνήεν αλλά ένας ουδέτερος ήχος που προέρχεται από κάποιο άτονο φωνήεν μέσα σε μια λέξη όπως η λέξη 'but' που ακούγεται /b @ t/ και όχι /b a t/.

Τα σύμφωνα

Τα σύμφωνα παράγονται λόγω τριβής που προκαλείται από τους στροβιλισμούς του αέρα. Η φύση του ήχου εξαρτάται από την περιοχή παραγωγής του και τη θέση της γλώσσας και των υπολοίπων μερών του φωνητικού συστήματος. Τέτοια σύμφωνα είναι τα /f/, /s/, /sh/.

Ένα άλλο είδος συμφώνου είναι το άηχο σύμφωνο, το οποίο δημιουργείται από την απότομη διακοπή του ρεύματος του αέρα από κάποιο διάφραγμα, όπως τα χείλη ή τα δόντια. Τότε ο αέρας συσσωρεύεται, η πίεση αυξάνει και όταν το διάφραγμα ανοίξει, έχουμε απότομη εκροή αέρα (όπως στα /p/, /k/).

Κατά τη διάρκεια του συνεχούς λόγου έχουμε και διαστήματα ησυχίας, αλλά όχι στην αρχή και στο τέλος μιας λέξης, όπως ίσως θα περιμέναμε, αλλά πριν τα άηχα σύμφωνα. Η διάρκεια αυτών των διαστημάτων είναι της τάξης των 30-50ms.

Τα ρινικά σύμφωνα παράγονται από την ρινική κοιλότητα ενώ η στοματική λειτουργεί ως συντονισμένο αντηχείο. Η ένταση του ήχου που βγαίνει είναι χαμηλότερη από τους άλλους λόγω του ότι παράγεται ευρύτερο φάσμα συχνοτήτων, ενώ παράλληλα η στοματική κοιλότητα απορροφά τον ήχο λόγω των μαλακών τοιχωμάτων.

Τα σύμφωνα χωρίζονται σε άφωνα τα οποία παράγουν μια τυχαία θορυβώδη ταλάντωση (λευκός θόρυβος) όπως /f/, /th/, /s/, /sh/ ή συνδυάζονται με παραγωγή φθόγγων όπως /v/, /z/, /dh/, /xh/.

Τα φωνήεντα

Η ταλάντωση των φωνητικών χορδών, με την παραγωγή των φωνηέντων παράγει κάποιες βασικές συχνότητες και τις αρμονικές τους οι οποίες είναι πολλαπλάσια της βασικής. Αν η βασική αρμονική είναι 100Hz τότε θα έχουμε αρμονικές 200, 300, 400, 500 Hz κλπ. Όμως, επειδή η κατασκευή του φωνητικού συστήματος έχει κάποιες ιδιοσυχνότητες, οι αρμονικές δεν ενισχύονται όλες το ίδιο. Συνήθως το φωνητικό σύστημα έχει δύο βασικές συχνότητες συντονισμού, f1 και f2.

Οι συχνότητες του κάθε φωνήεντος είναι αρκετά διαφορετικές από φωνήεν σε φωνήεν. Όμως για όλους τους ανθρώπους , οι συχνότητες συντονισμού για το ίδιο

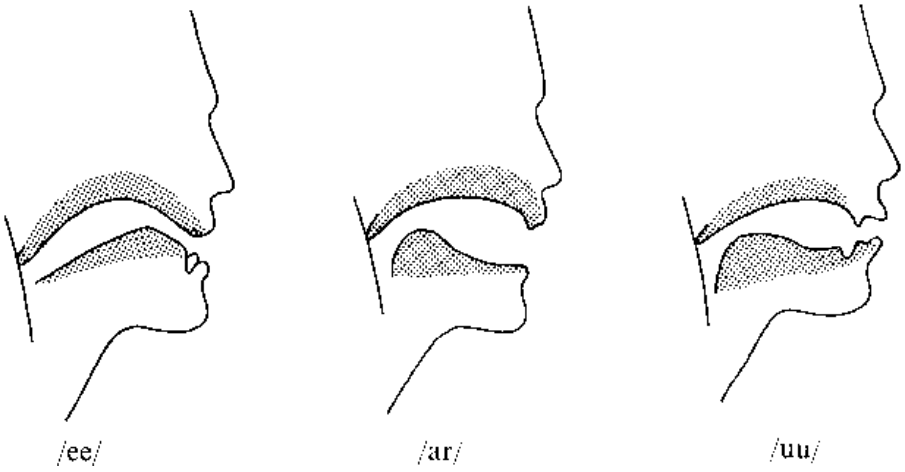
φωνήεν είναι περίπου οι ίδιες. Στο φωνήεν /ee/ όπως στη λέξη “he” η f1 είναι 300Hz και η f2 είναι 2100Hz . Η βασική συχνότητα μπορεί να διαφέρει από άνθρωπο σε άνθρωπο, από τη διάθεση του και τη προσωδία. Τελικά αυτό που κάνει τους ήχους να ξεχωρίζουν είναι το πλάτος και οι σχέσεις των συχνοτήτων συντονισμού.

Στον παρακάτω πίνακα 1.2 φαίνεται η αντιστοιχία φωνηέντων με τις συχνότητες συντονισμού του φωνητικού συστήματος.

Φωνήεν		f_1	f_2	f_3
/ee/	beat	280	2620	3380
/i/	bit	360	2220	2960
/e/	bet	600	2060	2840
/er/	bird	560	1480	2520
/ar/	father	740	1110	2640
/a/	hut	760	1370	2500
/u/	hood	480	740	2620
/uu/	loot	320	920	2200

Πίνακας 1.2 Οι βασικές συχνότητες συντονισμού για μερικά από τα φωνήεντα της Αγγλικής γλώσσας

Στην εικόνα 1.4 μπορούμε να δούμε πως περίπου είναι η εσωτερική δομή της στοματικής κοιλότητας την ώρα που αρθρώνονται οι φθόγγοι /ee/ όπως “ποίημα”, /ar/ όπως “αρχηγός” και /uu/ όπως “ούτε”.

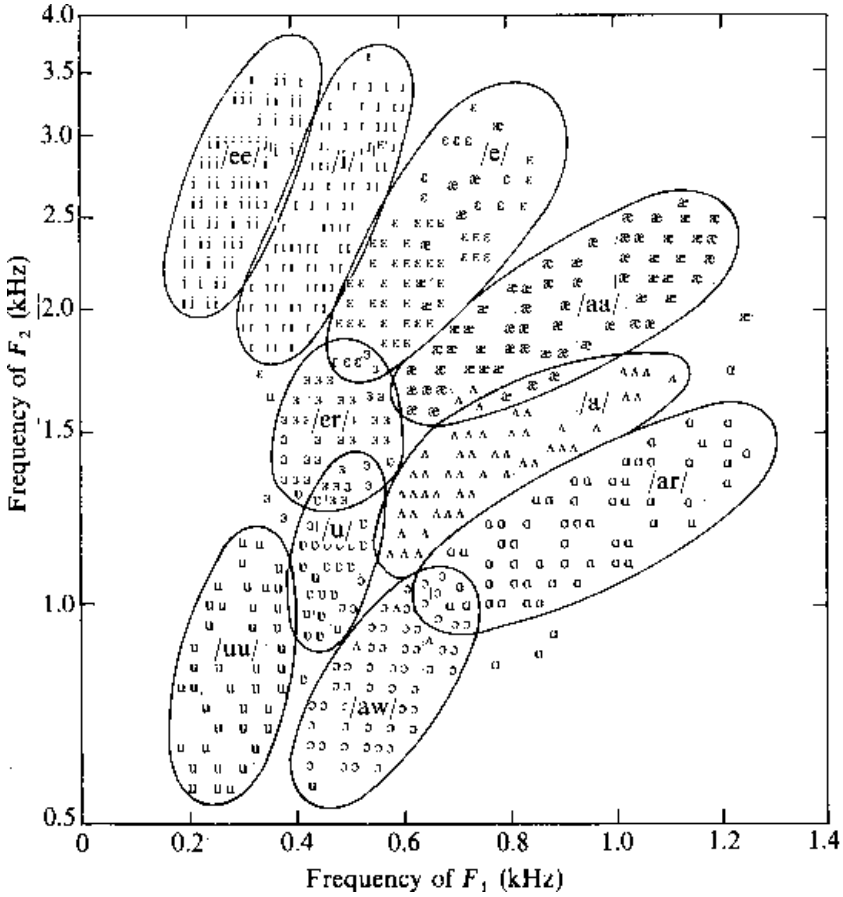


Εικόνα 1.4 Οι θέσεις που λαμβάνει το στόμα για την δημιουργία διαφόρων φθόγγων

Όταν παράγουμε το /ee/ τότε η γλώσσα κινείται μπροστά και προς τα πάνω, μειώνοντας τον όγκο της στοματικής κοιλότητας. Αυτό έχει ως αποτέλεσμα την παραγωγή δεύτερων και τρίτων αρμονικών στις συχνότητες συντονισμού. Αντίθετα όταν προφέρουμε το /ar/ η γλώσσα πηγαίνει προς τα πίσω και κάτω, ενώ το στόμα

ανοίγει, με αποτέλεσμα το μεγάλωμα της κοιλότητας και την παραγωγή υψηλότερων συχνοτήτων συντονισμού, που φτάνουν τα 750Hz με δεύτερη αρμονική τα 110Hz. Τέλος στο /uu/ η πρώτη και η δεύτερη αρμονική χαμηλώνουν, σηκώνοντας την γλώσσα προς τον ουρανίσκο και εκτείνοντας τα χείλη προς τα έξω, οπότε έχουμε επιμήκυνση της στοματικής κοιλότητας. Στην περίπτωση αυτή οι τρεις συχνότητες συντονισμού είναι 300, 900 και 2500Hz.

Στο σχήμα 1.5 φαίνεται ένα διάγραμμα των δύο βασικών συχνοτήτων συντονισμού με φωνήεντα από διάφορους ομιλητές. Βλέπουμε ότι τα σύνολα σπάνια αλληλεπικαλύπτονται και αυτό εξηγεί γιατί τα φωνήεντα είναι εύκολο να ξεχωρίσουν μεταξύ τους.



Εικόνα 1.5 Η κατανομή των βασικών συχνοτήτων για τα φωνήεντα σε διάφορους ομιλητές

Δίφθογγοι

Οι δίφθογγοι είναι συνδυασμός δύο φθόγγων. Η στοματική κοιλότητα αλλάζει σιγά σιγά ώστε να αλλάξει η βασική συχνότητα συντονισμού , οπότε έχουμε μια ομαλή μετάβαση αρμονικών.

Ημιφωνήεντα

Τα ημιφωνήεντα μοιάζουν με τα φωνήεντα, με τις χορδές να πάλλονται. Υπάρχουν δύο κατηγορίες ημιφωνηέντων τα μεταβατικά και τα υγρόληκτα. Τα μεταβατικά εμφανίζονται πριν ή ενδιάμεσα από φωνήεντα περνώντας από το ένα φωνήεν στο άλλο και έτσι μοιάζουν με τους δίφθογγους μόνο που η μετάβαση αυτή γίνεται πιο γρήγορα. Στα μεταβατικά οι στενώσεις της στοματικής οδού είναι μεγαλύτερες κατά
τη διάρκεια της μετάβασης. Οι αρθρώσεις αυτές καταλήγουν σε ασθενέστερες αλλά γρηγορότερες μεταβάσεις των formants. Τα υγρόληκτα είναι όμοια με τα μεταβατικά στον τρόπο που εκφωνούνται, στην ταχύτητα της κίνησης και στο βαθμό της στένωσης της στοματικής οδού αλλά διαφέρουν στον σχηματισμό της στένωσης από την γλώσσα. Η γλώσσα παίρνει τέτοιο σχήμα σαν να χαρίζει την στοματική οδό με αποτέλεσμα να δημιουργούνται αντί-αντηχήσεις (antiresonance).

Ρινικά σύμφωνα

Τα ρινικά σύμφωνα είναι η πλησιέστερη ομάδα συμφώνων στα φωνήεντα. Περιλαμβάνει τα σύμφωνα /m/ και /n/.

Πηγή: Όπως και με τα φωνήεντα έτσι και εδώ έχουμε φαινομενικά περιοδικούς παλμούς λόγω των ταλαντευόμενων φωνητικών χορδών.

Σύστημα: Η υπερώα είναι χαμηλωμένη και ο αέρας ρέει κυρίως μέσα από την ρινική κοιλότητα με την στοματική να είναι στενευμένη. Συνεπώς ο ήχος διαχέεται στα ρουθούνια. Παρατηρούμε ότι η στένωση στο /m/ γίνεται στα χείλη ενώ στο /n/ με την γλώσσα στα ούλα.

Φασματικό Περιεχόμενο: το φασματογράφημα ενός ρινικού συμφώνου κυριαρχείτε από την χαμηλή αντήχηση υψηλής έντασης της ρινικής κοιλότητας. Οι αντηχήσεις της ρινικής κοιλότητας έχουν μεγάλο εύρος διότι συμβαίνουν μεγάλες απώλειες καθώς ο αέρας ρέει δια μέσω της πολύπλοκης επιφάνειάς της εξασθενίζοντας γρήγορα την κρουστική της απόκριση. Η κλειστή στοματική οδός λειτουργεί σαν μια διακλάδωση που δημιουργεί τις δικές της αντηχήσεις. Οι αντηχήσεις αυτές απορροφούν ακουστική ενέργεια και συνεπώς είναι αντί-αντηχήσεις (μηδενικά) της φωνητικής οδού. Οι αντί-αντηχήσεις της στοματικής οδού εκτείνονται πέρα από τις αντηχήσεις τις ρινικής οδού με αποτέλεσμα η συνάρτηση μεταφοράς για τα ρινικά σύμφωνα να φέρει λίγη ενέργεια υψηλής συχνότητας. Έτσι για το /m/ υπάρχει ένα χαμηλό formant στα 250 Hz με λίγη ενέργεια πάνω από αυτήν. Ανάλογο πρότυπο έχει και το /n/.

Θορυβώδη (Fricatives)

Τα fricative σύμφωνα ταξινομούνται σε δύο κατηγορίες: τα ηχηρά (voiced) και τα άηχα (unvoiced) fricative.

Πηγή: Στην περίπτωση των άηχων, οι φωνητικές χορδές είναι χαλαρές και δεν ταλαντώνονται. Ο θόρυβος παράγεται από τη ‘βίαιη’ ροή αέρα σε μέσα από κάποιο στένωμα της στοματικής οδού, που είναι μικρότερο από αυτό στα φωνήεντα. Τα ηχηρά έχουν παρόμοια πηγή. Ωστόσο, συχνά οι φωνητικές χορδές ταλαντώνονται ταυτόχρονα με την παραγωγή θορύβου όπως στο fricative /z/ ή στο /v/ αντίθετα με το άηχο /f/.

Σύστημα: Η τοποθέτηση της στένωσης από την γλώσσα πίσω, στο κέντρο ή μπροστά στη στοματική οδό επηρεάζει ποιος fricative ήχος θα παραχθεί. Η στένωση διαχωρίζει τη στοματική οδό σε δύο κοιλότητες μπροστά και πίσω με τον ήχο να διαδίδεται από την μπροστινή. Στα ηχηρά fricative οι φωνητικές χορδές ταλαντώνονται και η περιοδική ροή από την γλωττίδα περνάει από την πίσω στοματική κοιλότητα προς την στένωση.

Φασματικό περιεχόμενο: Η μπροστινή κοιλότητα κυριαρχεί στον φασματικό σχεδιασμό του ήχου ενώ η πίσω δημιουργεί μηδενικά (αντί-αντηχήσεις) στη συνάρτηση μεταφοράς απορροφώντας ενέργεια. Επειδή η μπροστινή κοιλότητα είναι πιο μικρή της συνολικής στοματικής κοιλότητας και επειδή οι αντί-αντηχήσεις είναι σε μικρότερη συχνότητα από της αντηχήσεις της μπροστινής κοιλότητας, η συνάρτηση μεταφοράς που προκύπτει αποτελείται κυρίως από αντηχήσεις υψηλής συχνότητας που μεταβάλλονται με την θέση της στένωσης. Τα άηχα fricatives χαρακτηρίζονται από θορυβώδες φάσμα ενώ τα ηχηρά εμφανίζουν θόρυβο και αρμονικές μαζί. Η φασματική φύση του ήχου χαρακτηρίζεται από την θέση της στένωσης. Παράδειγμα με ένα /S/ η στένωση είναι στον ουρανίσκο ενώ στο /f/ στα χείλη. Έτσι το /S/ έχει υψιπερατό φάσμα που αντιστοιχεί σε μια άνω στενή στοματική κοιλότητα ενώ για το /f/ υπάρχει μια μπροστινή κοιλότητα οπότε έχουμε
ένα επίπεδο φάσμα. Τα ηχηρά fricatives χαρακτηρίζονται από την αρμονική δομή που οφείλεται στη ταλάντωση των χορδών. Στο φασματογράφημα εμφανίζεται ακόμη και η επιρροή από τα γειτονικά φωνήεντα στην μετατόπιση των formants.

Κυματομορφή: Στην κυματομορφή μπορούμε να παρατηρήσουμε σε ένα άηχο fricative τον θόρυβο ενώ σε ένα ηχηρό βλέπουμε θόρυβο και περιοδικότητα μαζί.

Έκκροτα (Plosives)

Όπως και τα fricatives, έτσι και τα plosives σύμφωνα διακρίνονται σε ηχηρά όπως το /b/ /d/ /g/ και άηχα όπως το /p/ /k/ /t/.

Πηγή και σύστημα: Τα άηχα plosives παράγονται δημιουργώντας πίεση πίσω από ένα κλείσιμο κάπου στην στοματική οδό και ελευθερώνοντάς την απότομα. Ανάλογα με το που είναι το κλείσιμο παράγεται και το αντίστοιχο plosive σύμφωνο. Για /p/ το κλείσιμο είναι στα χείλη, για το /t/ στα δόντια και για το /k/ στον ουρανίσκο. Παρόμοιος είναι ο τρόπος παραγωγής των ηχηρών με την μόνη διαφορά ότι οι φωνητικές χορδές πάλλονται. Καθ' όλη τη διάρκεια του κλεισίματος στα άηχα δεν παράγεται ήχος, στα ηχηρά όμως υπάρχει συχνά ένα μικρό ποσό ενέργειας χαμηλής συχνότητας που βγαίνει από τα τοιχώματα του λαιμού. Αυτό συμβαίνει όταν οι χορδές μπορούν να πάλλονται ακόμα και όταν η φωνητική οδός είναι κλειστή σε κάποιο σημείο. Οι ιδιότητες των ηχηρών plosives εξαρτώνται από το φωνήεν που ακολουθεί.

Φασματογράφημα και κυματομορφή: στα άηχα plosives παρατηρούμε και στο φασματογράφημα και στην κυματομορφή μια μικρή σιωπή, που φαίνεται στο φάσμα σαν κενό, που ακολουθείται από απότομο σπασμό και θόρυβο αναπνοής. Το φάσμα την στιγμή του σπασμού καθορίζεται από το σχήμα της στοματικής κοιλότητας μπροστά από το κλείσιμο. Οι τροχιές των formants καθώς μεταβαίνουμε από το plosive σε ηχηρή κατάσταση απεικονίζουν την μεταβολή της φωνητικής οδού και πολλές φορές από τις τροχιές αυτές μπορούμε να καταλάβουμε αν το plosive είναι ηχηρό ή άηχο.

Χαρακτηριστικά φωνής

Η μορφή της φωνής εξαρτάται από πολλούς παράγοντες όπως η συχνότητα διέγερσης των φωνητικών χορδών, το μέγεθος του λάρυγγα, το μήκος των φωνητικών χορδών και τον τονισμό των συλλαβών. Επίσης, η ένταση της ομιλίας εξαρτάται από την συναισθηματική κατάσταση που βρίσκεται ο ομιλητής, από το πόσο μακριά θέλουμε να μεταδοθεί το μήνυμα, και από το πόσο θορυβώδες είναι το περιβάλλον. Ακόμη, κατά τη διάρκεια της ομιλίας δημιουργούνται διάφορα φωνητικά φαινόμενα τα οποία επηρεάζουν το τελικό αποτέλεσμα του λόγου.

Το πρώτο από αυτά τα φαινόμενα είναι η *προσαρμογή*. Αυτή προκύπτει επειδή το φωνητικό σύστημα κατά τη διάρκεια εκφώνησης κάποιου φθόγγου, προετοιμάζει τη θέση των οργάνων παραγωγής ήχου για τον επόμενο. Όταν μιλάμε γρήγορα τότε η προσαρμογή είναι πιο έντονη, γιατί η γλώσσα δεν πηγαίνει στις θέσεις που πρέπει όπως όταν μιλάμε αργά. Σε οριακές περιπτώσεις η προσαρμογή μπορεί να επιφέρει *αφομοίωση.* Σ' αυτή την περίπτωση ο ήχος του φθόγγου αλλάζει και παίρνει πολλά χαρακτηριστικά από τον επόμενο με αποτέλεσμα την ενοποίηση του όπως για παράδειγμα στη λέξη “link” όπου το /n/ γίνεται /ng/.

Όταν δύο διαφορετικά όργανα κινούνται ταυτόχρονα για δύο διαφορετικούς φθόγγους, τότε έχουμε *συνδυασμούς άρθρωσης*. Το αποτέλεσμα της χρήσης της προσαρμογής και των συνδυασμών άρθρωσης είναι η καλύτερη ποιότητα του ήχου.

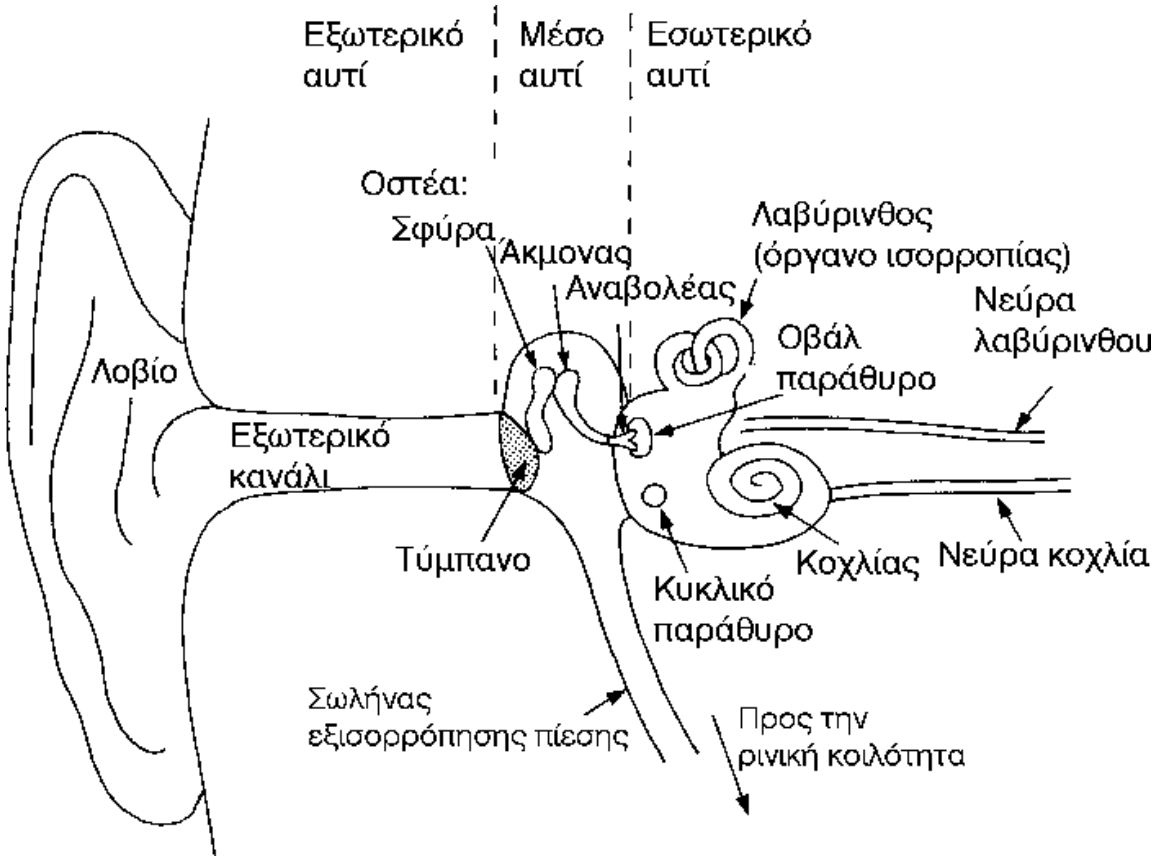
Δεν πρέπει όμως να βλέπουμε τον ήχο της ομιλίας απομονωμένα σαν σκέτους φθόγγους, αφού αυτοί αποτελούν την βασική μονάδα για την υλοποίηση των λέξεων, των προτάσεων και των φράσεων. Η ομιλία έχει και άλλα χαρακτηριστικά τα οποία εξαρτώνται από το νόημα της φράσης. Έτσι έχουμε το χαρακτηριστικό της έντασης. Η ένταση σημαίνει την αύξηση της έντασης, του τόνου, της διάρκειας και του τονισμού κάποιων συλλαβών. Ο τονισμός είναι ένα άλλο χαρακτηριστικό της φωνής. Σ' αυτή την περίπτωση έχουμε αλλαγή της βασικής συχνότητας κάνοντας τον τόνο της φωνής να ανεβαίνει ή να κατεβαίνει δίνοντας πρόσθετη σημασία, όπως για παράδειγμα στις ερωτήσεις. Τέλος υπάρχει και το χαρακτηριστικό της διάρκειας, όπου δύο προτάσεις που διαφέρουν σε νόημα ακούγονται το ίδιο, και ο μόνος τρόπος να διαχωριστούν είναι κάποια παύση στη μέση της πρότασης όπως “an aim” και “a name”. Ένα άλλο χαρακτηριστικό είναι η προφορά του κάθε ομιλητή ανάλογα με την καταγωγή του και την περιοχή όπου μεγάλωσε.

Ακουστική

Για να κατανοήσουμε τον τρόπο που ο άνθρωπος αντιλαμβάνεται την ομιλία και την αναγνωρίζει, είναι χρήσιμο να ερευνήσουμε όχι μόνο τον τρόπο παραγωγής αλλά και τον τρόπο λήψης ήχων από το ακουστικό σύστημα του ανθρώπου.

Το αυτί

Το αυτί μπορεί να χωριστεί σε τρία βασικά κομμάτια. Το πρώτο είναι το εξωτερικό αυτί το οποίο προσφέρει προστασία στο μεσαίο και εσωτερικό αυτί. Επίσης είναι η διάταξή του που επιτρέπει την μερική κατευθυντικότητα των λαμβανόμενων ήχων από την μπροστινή μεριά του κεφαλιού. Πρέπει να σημειωθεί βέβαια ότι οι μηχανισμοί εντοπισμού της ηχητικής πηγής είναι με βάση την ένταση και το συγχρονισμό των δυο αυτιών.



Εικόνα 1.6 Το αυτί και τα όργανα της ακοής

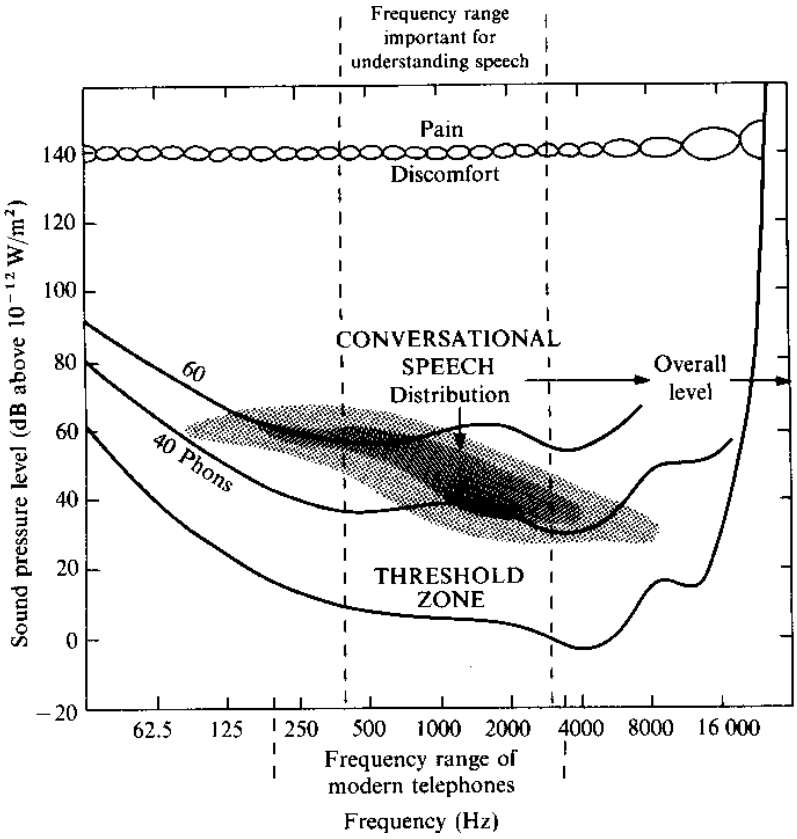
Το ακουστικό νεύρο

Το ακουστικό νεύρο χρησιμοποιεί τρεις τρόπους μεταφοράς της πληροφορίας προς τον εγκέφαλο. Στέλνει ριπές σήματος στον εγκέφαλο όσο υπάρχει η απουσία σχετικού σήματος. Το κάθε ακουστικό νεύρο έχει επιλεκτικότητα, δηλαδή καλύτερη απόκριση σε ορισμένες συχνότητες. Τέλος το νεύρο στέλνει ηλεκτρικά ερεθίσματα κατά ριπές που παρακολουθούν την φάση του βασικού σήματος που διεγείρει την μεμβράνη. Στο διάγραμμα της συχνότητας σε σχέση με την πίεση του ήχου (εικόνα 1.7) παρατηρούμε ότι η ένταση που χρειάζεται για να ενεργοποιηθεί ένας νευρώνας είναι μικρή για κάποια ορισμένη χαρακτηριστική συχνότητα. Η καμπύλη από τις χαμηλότερες συχνότητες προς την χαρακτηριστική είναι λιγότερο απότομη από ότι

είναι στην χαρακτηριστική μέχρι τις υψηλότερες συχνότητες. Ο ρυθμός αποφόρτισης του νεύρου και η ένταση του ήχου μέχρι ένα σημείο είναι ανάλογα μεγέθη. Η ενεργοποίηση του νευρώνα αρχίζει από τα 20db περίπου και ο κορεσμός είναι στα 50db. Οι νευρώνες παρακολουθούν τη φάση στην περιοχή μεταξύ 4 και 5 KHz. Σε μεγαλύτερες συχνότητες δεν υπάρχει παρακολούθηση της φάσης.

Όταν έχουμε παρουσία δυο τόνων η πρώτη συχνότητα μπορεί να μειωθεί ως προς την δεύτερη. Αυτό το φαινόμενο καλείται απόκρυψη (two tone suppression). Αν η δεύτερη συχνότητα είναι πολύ κοντά στην πρώτη και μέσα στην περιοχή των ρίπων, τότε αυτή ενισχύεται. Αν είναι έκτος της περιοχής αυτής, καταπνίγεται. Αυτό συμβαίνει λόγω της μη γραμμικότητας της μεμβράνης. Σε πειράματα που έγιναν μετρήθηκε μια σύνθετη συχνότητα $2f_1$-f_2 όπου f_1 η μια συχνότητα, f_2 η δεύτερη και, $f_1<f_2$.

Το ανθρώπινο αυτί δεν είναι το ίδιο ευαίσθητο σε όλη την περιοχή των συχνοτήτων. Η μεγαλύτερη ευαισθησία είναι μεταξύ των 1000Hz και 4000Hz. Οι υψηλότερος και χαμηλότερες συχνότητες χρειάζονται μεγαλύτερη ενέργεια για να γίνουν αντιληπτές. Στην παρακάτω εικόνα 1.7 βλέπουμε την απόκριση συχνότητας σε σχέση με την ένταση με 0 db σε πίεση 10^{-12} W/m^2.



Εικόνα 1.7 Η απόκριση του αυτιού και τα όρια ακουστότητας

Η χαμηλότερη καμπύλη δείχνει την ελάχιστη ένταση που απαιτείται για να γίνει αντιληπτός ο ήχος. Αυτή είναι μια τυπική καμπύλη και διαφέρει ανάλογα με τον άνθρωπο μέχρι και 20db. Η ευαισθησία στις υψηλές συχνότητες μειώνεται με την

ηλικία. Οι άλλες δυο καμπύλες δείχνουν τα σημεία που έχουν την ίδια ένταση για το αυτί σε phon. Το phon μετριέται σε db με πίεση πάνω από $10^{-12}W/m^2$ στα 1000Hz. Τέλος η επιφάνεια στη μέση δείχνει την κατανομή της ανθρώπινης ομιλίας, το μεγαλύτερο ποσοστό της οποίας πέφτει μέσα στην ευαίσθητη περιοχή του αυτιού. Μικρές αλλαγές στην πίεση (0.5-2db) μπορούν να γίνουν αντιληπτές λόγω του εύρους που μπορεί να μεταβάλλεται ο ρυθμός των ρίπων στα νεύρα.

Διάκριση των τόνων

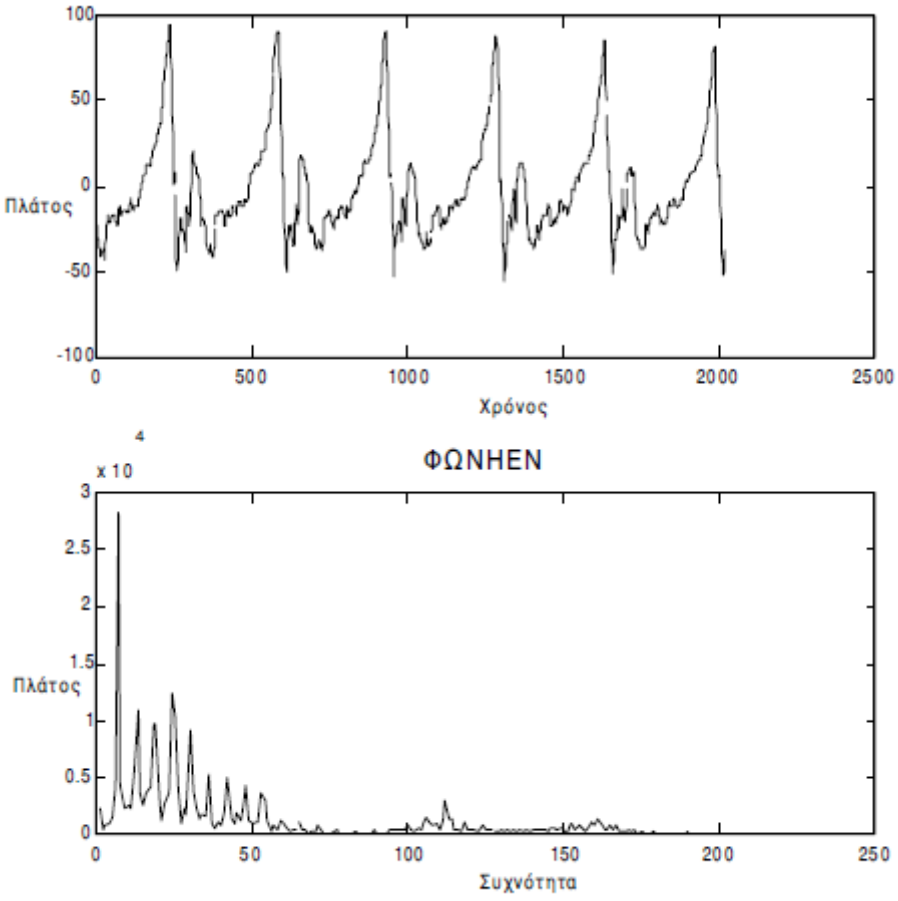
Η ανίχνευση των τόνων μας επιτρέπει να κατατάξουμε τους ήχους στη μουσική κλίμακα. Όταν το αυτί ακούει έναν καθαρό τόνο τότε αντιλαμβανόμαστε την συχνότητά του. Όταν ακούμε έναν σύνθετο τόνο που περιέχει πολλές διαφορετικές συχνότητες τότε ακούμε τη βασική αρμονική. Υπάρχουν δύο θεωρίες που εξηγούν τον τρόπο διάκρισης των τόνων. Η πρώτη συσχετίζει τον τόνο με τη θέση της μέγιστης διέγερσης της βασικής μεμβράνης. Η δεύτερη συσχετίζει τον τόνο με τις χρονικές παραλλαγές των νευρωνικών παλμών. Καμιά από τις δύο θεωρίες δεν εξηγεί πλήρως την λειτουργία του αυτιού. Ίσως η λύση να είναι ο συνδυασμός τους.

Είναι σημαντικό για τους ανθρώπους να μπορούν να διαχωρίσουν διαφορετικές συχνότητες, όπως για παράδειγμα στην ομιλία. Κάθε φωνήεν έχει πολλές συχνότητες που πρέπει να γίνουν αντιληπτές ταυτόχρονα ώστε να γίνει η αναγνώριση. Ο Fletcher (1940) πρώτος διατύπωσε τη θεωρία ότι οι διάφορες συχνότητες γίνονται αντιληπτές από μία σειρά αλληλοεπικαλυπτόμενων ζωνοπερατών φίλτρων που η βασική τους συχνότητα αλλάζει σαρώνοντας όλη την ακουστική περιοχή. Αυτή η θεωρία έχει επιβεβαιωθεί με διάφορα πειράματα που έγιναν. Η δυνατότητα διαχωρισμού δύο γειτονικών συχνοτήτων που είναι πολύ κοντά μεταξύ τους, περιορίζεται από το πλάτος του ενός εκ των δύο ακουστικών φίλτρων, το οποίο καλείται κρίσιμο πλάτος.

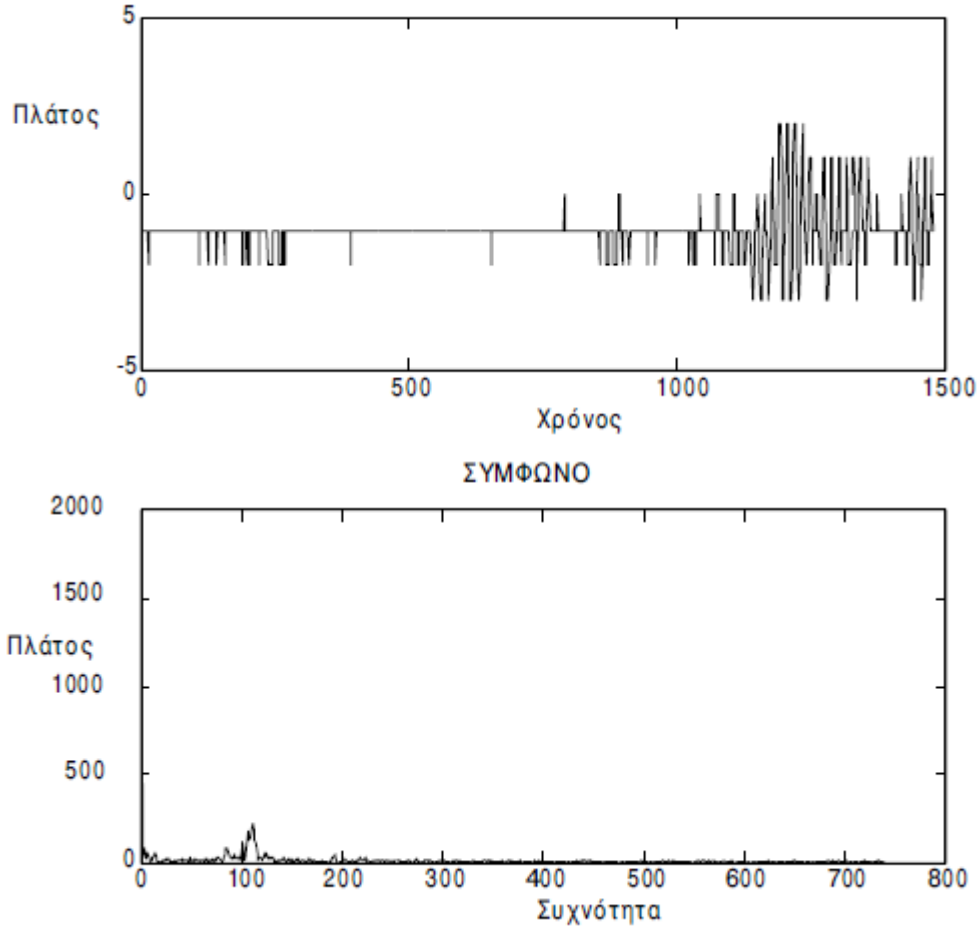
Η επιλεκτικότητα των συχνοτήτων συσχετίζεται και με την δυνατότητα της απόκρυψης συχνότητας. Αυτό σημαίνει ότι ένας ήχος παύει να ακούγεται μόλις εμφανιστεί ένας άλλος ήχος, όπως γίνεται στο αυτοκίνητο όταν ανοίγουμε το ραδιόφωνο, όπου δεν ακούμε τον θόρυβο του αυτοκινήτου με την ίδια ένταση. Η απόκρυψη αυτή είναι πολύ πιο αποτελεσματική όταν οι δύο συχνότητες είναι πολύ κοντά μεταξύ τους. Ο Fletcher έκανε πειράματα για την απόκρυψη των συχνοτήτων. Χρησιμοποίησε μια γεννήτρια ημιτόνου και μία γεννήτρια λευκού θορύβου με μεταβλητό εύρος. Παρατήρησε λοιπόν ότι όσο αύξανε το εύρος του θορύβου (με σταθερή την ισχύ) τόσο αύξανε και το επίπεδο της έντασης που χρειαζόταν για να γίνει αντιληπτή η βασική συχνότητα του ημιτόνου. Αυτό συνέβαινε μέχρι ενός σημείου, γιατί κατόπιν όσο αυξανόταν το εύρος του θορύβου, δεν υπήρχε καμία αλλαγή στο επίπεδο της έντασης που ήταν απαραίτητη για την αναγνώριση του ήχου. Αυτό το εύρος το καλούμε κρίσιμο. Το κρίσιμο εύρος είναι περίπου το 10-20% της βασικής συχνότητας. Αυτή η περίπτωση απόκρυψης είναι η ταυτόχρονη. Υπάρχει και η δυνατότητα προ-απόκρυψης όπου ο ήχος μπορεί να υποκρυφθεί από έναν προγενέστερο του ή αντίστοιχα για την μετά-απόκρυψη από έναν μεταγενέστερο.

Χαρακτηριστικά ομιλίας

Το σήμα της ομιλίας είναι ένα πολυδιάστατο σήμα όπου το πλάτος, η συχνότητα και η φάση μεταβάλλονται συνεχώς. Αυτό καθιστά την αναγνώριση ομιλίας από τις μηχανές ένα ιδιαίτερα δύσκολο πρόβλημα. Στην παρακάτω εικόνα 1.8 βλέπουμε ένα φωνήεν στο πεδίο του χρόνου, δηλαδή όπως θα το βλέπαμε σ' έναν παλμογράφο. Παρατηρείστε την ένταση του σήματος και τη μορφή του. Στα φωνήεντα έχουμε μεγάλο πλάτος και περιοδικότητα κάποιας βασικής συχνότητας. Στο πεδίο των συχνοτήτων αυτό σημαίνει μεγάλη ισχύ στις χαμηλές συχνότητες με κορυφές στις βασικές συχνότητες, και μειωμένη ισχύ στις υψηλές. Αντίθετα ένα σύμφωνο όπως φαίνεται στην εικόνα 1.9 έχει μικρό πλάτος και διασχίζει πολλές φορές τον άξονα του μηδενός (πολλές εναλλαγές προσήμου). Στο πεδίο των συχνοτήτων φαίνεται η χαμηλή ισχύς του σήματος, ενώ αντίθετα είναι σαφές ότι το εύρος των συχνοτήτων είναι πολύ μεγάλο. Παρατηρούμε ότι το φάσμα μοιάζει με λευκό θόρυβο (όπως ο ήχος που παράγεται από τον αέρα που βγαίνει μέσα από τα δόντια στο "σ"), ενώ δεν υπάρχει κάποια συγκεκριμένη συχνότητα που να συγκεντρώνει μεγάλη ενέργεια.

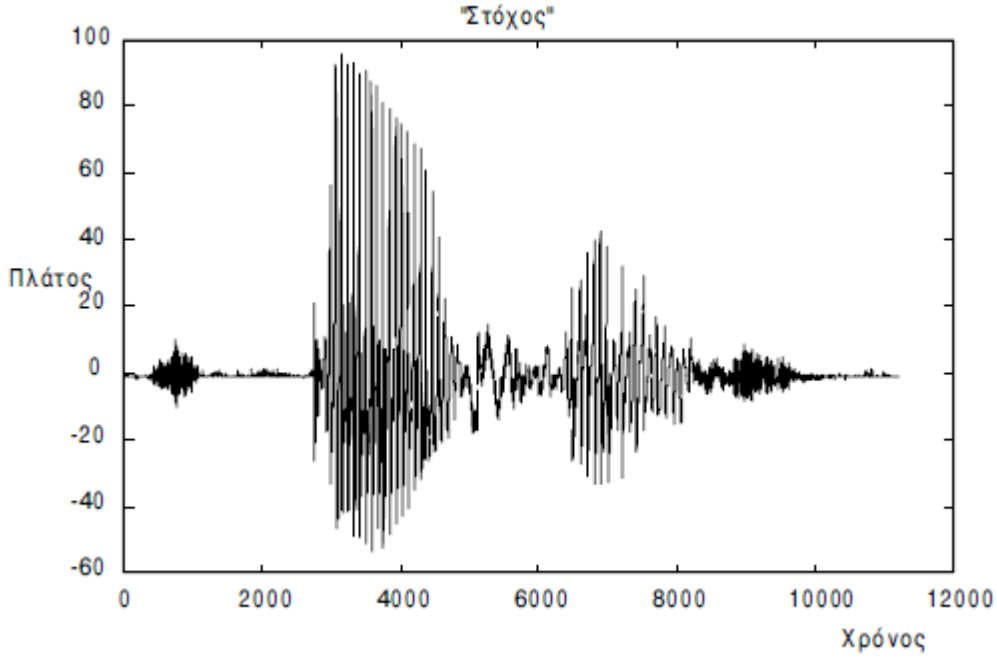


Εικόνα 1.8 Η κυματομορφή ενός φωνήεντος και το φάσμα του



Εικόνα 1.9 Η κυματομορφή ενός συμφώνου και το φάσμα του

Το πρώτο πρόβλημα που αντιμετωπίζει κανείς στην προσπάθεια κατασκευής ενός συστήματος αναγνώρισης ομιλίας είναι η αναγνώριση της αρχής και του τέλους μιας λέξης. Η αρχή μιας λέξης μπορεί να βρεθεί με την εύρεση της ενέργειας και τον αριθμό των εναλλαγών γύρω από τον άξονα του μηδενός. Όταν οι τιμές αυτές αυξηθούν πάνω από κάποιο όριο, έχουμε εντοπίσει την αρχή της λέξης. Το τέλος μιας λέξης είναι πιο δύσκολο να βρεθεί. Στην εικόνα 1.10 βλέπουμε τη λέξη "στόχος". Παρατηρείστε τα σύμφωνα "στ" όπου έχουμε κυματομορφή παρόμοια με θόρυβο, και ένα μεγάλο περιθώριο ησυχίας στο "τ". Έτσι φαίνεται ότι δεν μπορούμε να προσδιορίσουμε το τέλος μιας λέξης απλώς και μόνο περιμένοντας τη σιγή, αλλά πρέπει να ληφθούν υπ' όψιν και άλλοι παράγοντες. Ένα άλλο πρόβλημα είναι το ότι ο ομιλητής δεν μιλάει πάντα με την ίδια ταχύτητα αλλά μπορεί να εκφράσει την ίδια λέξη με πολλές διαφορετικές ταχύτητες ανάλογα με την περίπτωση. Ένα σύστημα αναγνώρισης θα πρέπει λοιπόν να λάβει και την ταχύτητα εκφώνησης υπ' όψιν του.



Εικόνα 1.10 Η κυματομορφή της λέξης ‘‘στόχος’’

Ένα από τα πολλά μοντέλα αναγνώρισης της ομιλίας που έχουν προταθεί αναλύει τη διαδικασία σε διάφορα βήματα. Πρώτα ο βασικός ακουστικός μηχανισμός απομονώνει τα βασικά στοιχεία της ομιλίας. Κατόπιν τα στοιχεία αυτά φιλτράρονται για να εντοπιστούν οι βασικές συχνότητες. Το επόμενο βήμα εντοπίζει τα χαρακτηριστικά των φθόγγων. Τέλος γίνεται τμηματική και λεξική ανάλυση. Βλέπουμε δηλαδή μια μεθοδολογία από κάτω προς τα επάνω (down-top) όπου σε κάθε βήμα επεξεργαζόμαστε σε πιο υψηλό επίπεδο την πληροφορία ώστε στο τέλος να καταλήξουμε σε λέξεις.

Διαδικασίες ανάλυσης φωνής

Για την ανάλυση και την επεξεργασία της φωνής χρησιμοποιούνται δυο προσεγγίσεις. Η πρώτη προσέγγιση χρησιμοποιεί μεθόδους που εξάγουν συμπεράσματα αναλύοντας τη φωνή στο πεδίο του χρόνου, ενώ η δεύτερη προσέγγιση αναλύει τα δεδομένα της φωνής στο πεδίο των συχνοτήτων. Αν κάνουμε μια αντίστοιχη διαστρωμάτωση στην επεξεργασία όπως ο άνθρωπος τότε η επεξεργασία του ήχου έχει τα εξής στάδια:

1. Επίπεδο πρωτογενούς επεξεργασίας σήματος (prefiltering, αρχή-τέλος λέξης κλπ)
2. Επίπεδο φθόγγου (διαχωρισμός φθόγγων, δυναμικός προγραμματισμός)
3. Επίπεδο λέξης (αναγνώριση, νοηματική επαλήθευση)

Με αυτή τη μεθοδολογία είναι φανερό το είδος της επεξεργασίας σε κάθε βήμα. Στο πρώτο επίπεδο η επεξεργασία γίνεται στο πεδίο του χρόνου. Στο δεύτερο επίπεδο γίνεται κυρίως στο πεδίο των συχνοτήτων, ενώ στο τελευταίο επίπεδο η ανάλυση –

επεξεργασία ξεφεύγει από τα στενά όρια της επεξεργασίας σήματος και μπορεί να είναι ένα στατιστικό σύστημα, ένα νευρωτικό δίκτυο ή ένα σύστημα ασαφούς λογικής.

Στο πεδίο του χρόνου μπορούμε να έχουμε σταθερές επεξεργασίες λόγο του ότι ο ήχος παρουσιάζεται στον επεξεργαστή ως μια ακολουθία αριθμών, όπου μια σύντομη επεξεργασία από δείγμα σε δείγμα αρκεί για να βγουν τα πρώτα συμπεράσματα.

Οι βασικές μέθοδοι επεξεργασίας συνεχούς ροής στο πεδίο του χρόνου είναι:

1. Μέσο πλάτος
2. Κατανομή πλάτους
3. Μέση ενέργεια
4. Ρυθμός διάβασης από το μηδέν

Μέσο πλάτος

Το μέσο πλάτος είναι ίσως ο ευκολότερος τρόπος να καταλάβει κανείς την ύπαρξη μιας λέξης ή αν υπάρχει φωνήεν ή σύμφωνο σε κάποια χρονική περιοχή. Το μέσο πλάτος υπολογίζεται σε ένα παράθυρο από τα τελευταία N δείγματα. Η σχέση είναι:

$$M(n) = \frac{1}{N} \sum_{m=n-N+1}^{n} |x(m)|$$

Όπου x(m) είναι το τρέχον δείγμα από μια απειροστή ακολουθία, ενώ το μέσο πλάτος είναι ένα παράθυρο στο χρόνο της ακολουθίας μήκους N.

Μια εναλλακτική μέθοδος υπολογισμού είναι η συνέλιξη μιας συνάρτησης παραθύρου με την απειροστή ακολουθία των δειγμάτων έτσι ώστε:

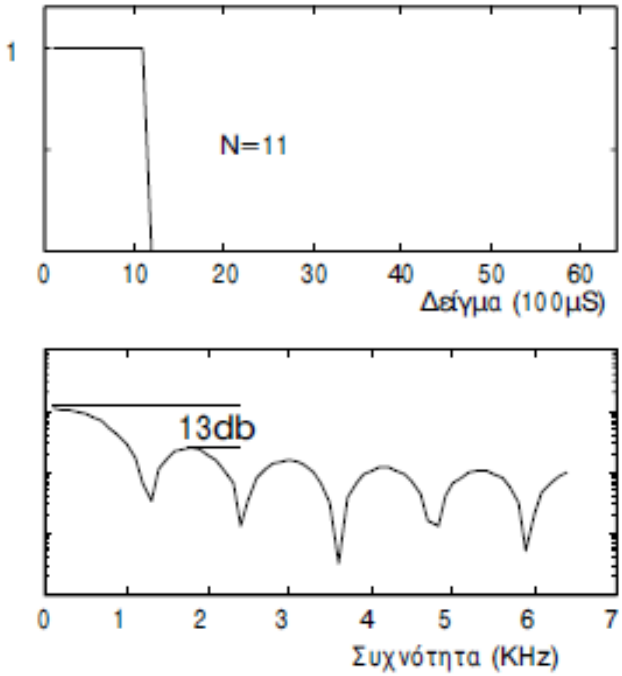
$$M(n) = \frac{1}{N} \sum_{m=-\infty}^{\infty} |x(m)| \cdot w(n-m)$$

Λόγω του ότι η συνέλιξη στο πεδίο του χρόνου ισοδυναμεί με πολλαπλασιασμό στο πεδίο των συχνοτήτων το μέσο πλάτος πρακτικά αντιστοιχεί με ένα φίλτρο με απόκριση w(). Τελικά η έξοδος της διαδικασίας θα είναι ανάλογη με ένα βαθυπερατό φίλτρο. Αυτό φαίνεται εύκολα αν υποθέσουμε ότι η w(n) γράφεται

$$w(n) = \begin{cases} 1, & 0 \le n \le N-1 \\ 0, & n < 0, n > N-1 \end{cases}$$

που ισοδυναμεί με ένα τετράγωνο παράθυρο.

Η φασματική απόκριση ενός τετραγωνικού παραθύρου μήκους N=11 με συχνότητα δειγματοληψίας 10 KSamples/sec φαίνεται στην εικόνα 1.11



Εικόνα 1.11 Τετραγωνικό παράθυρο μήκους 11 και η απόκρισή του

Είναι φανερό ότι η απόκρισή του είναι ένα βαθυπερατό φίλτρο με συχνότητα αποκοπής τα 700Hz. Αυτό αντιστοιχεί σε υποδειγματοληψία αφού ο ρυθμός δειγματοληψίας από 10KSamples/sec μπορεί να κατέβει στα 1.4 KSamples/sec σύμφωνα με το θεώρημα της δειγματοληψίας. Το τετραγωνικό παράθυρο δεν είναι το καλύτερο που υπάρχει αλλά είναι πολύ εύκολο στην υλοποίηση του. Υπάρχουν παράθυρα όπως πχ. Hamming όπου οι λοβοί στο πεδίο των συχνοτήτων έχουν πολύ μικρότερη ισχύ από αυτούς του τετραγωνικού παραθύρου.

Το μέσο πλάτος μπορεί να χρησιμοποιηθεί σε ένα σύστημα αυτομάτου ελέγχου της ενίσχυσης (AGC) ώστε να υπάρχει σταθερό πλάτος στα επόμενα στάδια της επεξεργασίας, όπως για παράδειγμα όταν το μικρόφωνο είναι σταθερό και ο ομιλητής κινείται μέσα στο χώρο.

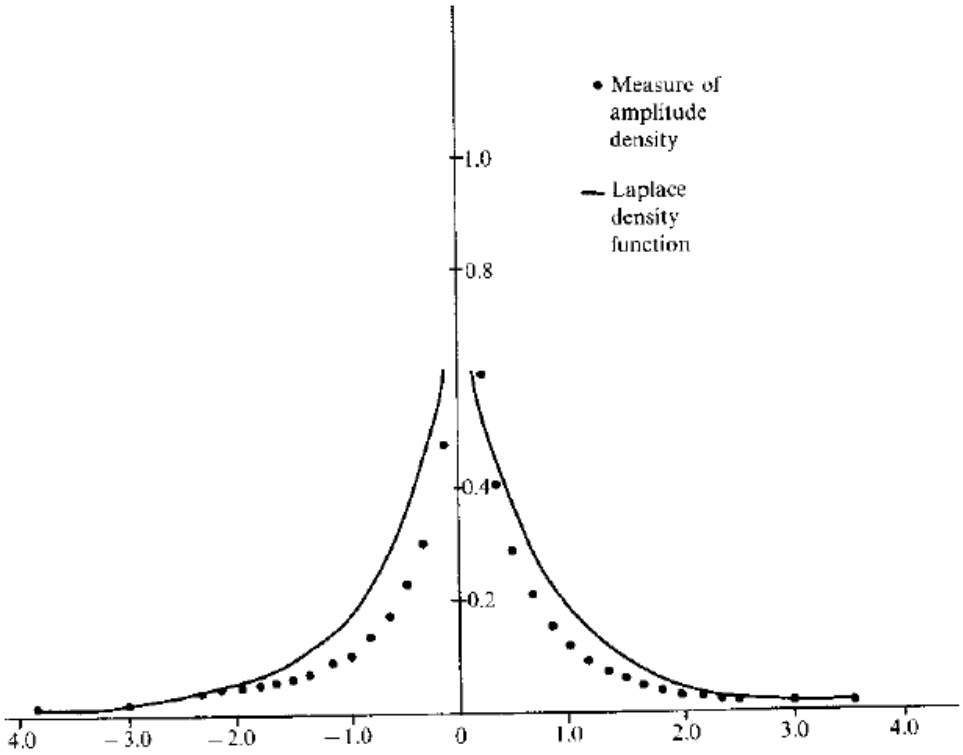
Κατανομή πλάτους

Το μέσο πλάτος μας δίνει τα χαρακτηριστικά της φωνής για μικρές χρονικές περιόδους. Σε περίπτωση που θέλουμε να γνωρίζουμε τα χαρακτηριστικά της φωνής που παράγονται σε μεγάλες χρονικές περιόδους, δεν αρκεί η χρήση του μέσου πλάτους. Σε αυτές τις περιπτώσεις χρησιμοποιείται η κατανομή πλάτους. Η κατανομή πλάτους είναι η στατιστική κατανομή του πλάτους των δειγμάτων. Η κατανομή των πιθανοτήτων μιας σειράς δειγμάτων μπορεί να υπολογιστεί από την αναλογία p(x) των δειγμάτων εισόδου x_i που βρίσκονται στο διάστημα $x < x_i < x+dx$. Οι τιμές των δειγμάτων της ομιλίας βρίσκονται στο διάστημα από $x = -\infty$ ως $x = +\infty$ έχουμε ότι:

$$\int_{-\infty}^{+\infty} p(x)dx = 1$$

Το άνω όριο της περιοχής τιμών των πιθανοτήτων στην κατανομή είναι ίσο με τη μονάδα.

Στην παρακάτω εικόνα 1.12 έχουμε τη γραφική απεικόνιση της κατανομής των πιθανοτήτων από ομιλία μήκους 12 δευτερολέπτων, με δειγματοληψία 16KSamples/sec και διακριτικότητα 14 bits. Το δείγμα είναι οι αριθμοί από μηδέν ως εννέα από μια γυναίκα και έναν άνδρα.



Εικόνα 1.12 Η κατανομή πλάτους μερικών λέξεων και η κατανομή Laplace

Οι τιμές των δειγμάτων έχουν κανονικοποιηθεί με βάση την rms τιμή του συνολικού δείγματος δηλαδή ένα δείγμα με πλάτος ίσο με την ολική rms τιμή έχει τιμή 1.0. Εκτός από τα δείγματα στο γράφημα φαίνεται και η συνάρτηση πυκνότητας Laplace

$$p(x) = \frac{1}{\sqrt{2\sigma_x}} \exp\left(-\frac{\sqrt{2|x|}}{\sigma_x}\right)$$

Η συγκεκριμένη συνάρτηση έχει το πλεονέκτημα ότι σχετίζεται άμεσα με το πλάτος rms του σήματος.

Η κατανομή του πλάτους είναι χρήσιμη όταν πρέπει να γίνει κλιμάκωση του αναλογικού σήματος που δειγματοληπτείται από ένα γραμμικό μετατροπέα A/D, βοηθώντας μας να επιλέξουμε τις τιμές εξόδου που χρησιμοποιούνται περισσότερο και να απορρίψουμε αυτές που συμμετέχουν λιγότερο.

Μέση ενέργεια

Η μέση ενέργεια είναι ο μέσος όρος των στιγμιαίων ενεργειών των δειγμάτων για κάποιο ορισμένο χρονικό διάστημα. Η ενέργεια σε κάποιο χρονικό διάστημα ορίζεται:

$$E(n) = \frac{1}{N} \sum_{m=-\infty}^{+\infty} [x(m)w(n-m)]^2$$

Όπου w() είναι η κατάλληλη συνάρτηση παραθύρου για να απομονώσει τα δείγματα που αντιστοιχούν στην περιοχή που μας ενδιαφέρει (η μέση ενέργεια γύρω από το δείγμα m και σε απόσταση n δειγμάτων).

Η παραπάνω εξίσωση μπορεί να γραφεί και ως:

$$E(n) = \frac{1}{N} \sum_{m=-\infty}^{+\infty} x^2(m)h(n-m)$$

με

$$h(n) = \frac{1}{N} w(\quad)^2$$

Η μέση ενέργεια περιορισμένης χρονικής διάρκειας από την παραπάνω σχέση, προκύπτει ότι μπορεί να υπολογιστεί αν φιλτράρουμε το τετράγωνο της εισόδου με τη συνάρτηση h().

Η μέση ενέργεια όπως φαίνεται και στη συνοπτική εικόνα 1.13 είναι σε μορφή παρόμοια με το μέσο πλάτος, αλλά έχει τονίσει τις διαφορές μεταξύ των φωνηέντων και των συμφώνων. Αυτός είναι και ο λόγος που η μέση ενέργεια προτιμάται για την ανίχνευση της ύπαρξης μιας λέξης. Συνήθως χρησιμοποιείται όμως σε συνδυασμό με το ρυθμό διάβασης από το μηδέν.

Ρυθμός διάβασης από το μηδέν

Η σημαντικότερη διαφορά μεταξύ των φωνηέντων και των συμφώνων είναι ο ρυθμός αλλαγής προσήμου. Τα σύμφωνα επειδή η κυματομορφή τους μοιάζει με λευκό θόρυβο με κατανομή Gauss, παρουσιάζουν πολλές εναλλαγές στο πρόσημο των δειγμάτων τους. Ο ρυθμός εναλλαγής του προσήμου για ένα φωνήεν είναι 0.5 εναλλαγές/ms ενώ για ένα σύμφωνο 3 εναλλαγές/ms. Ο ρυθμός αυτός μπορεί να

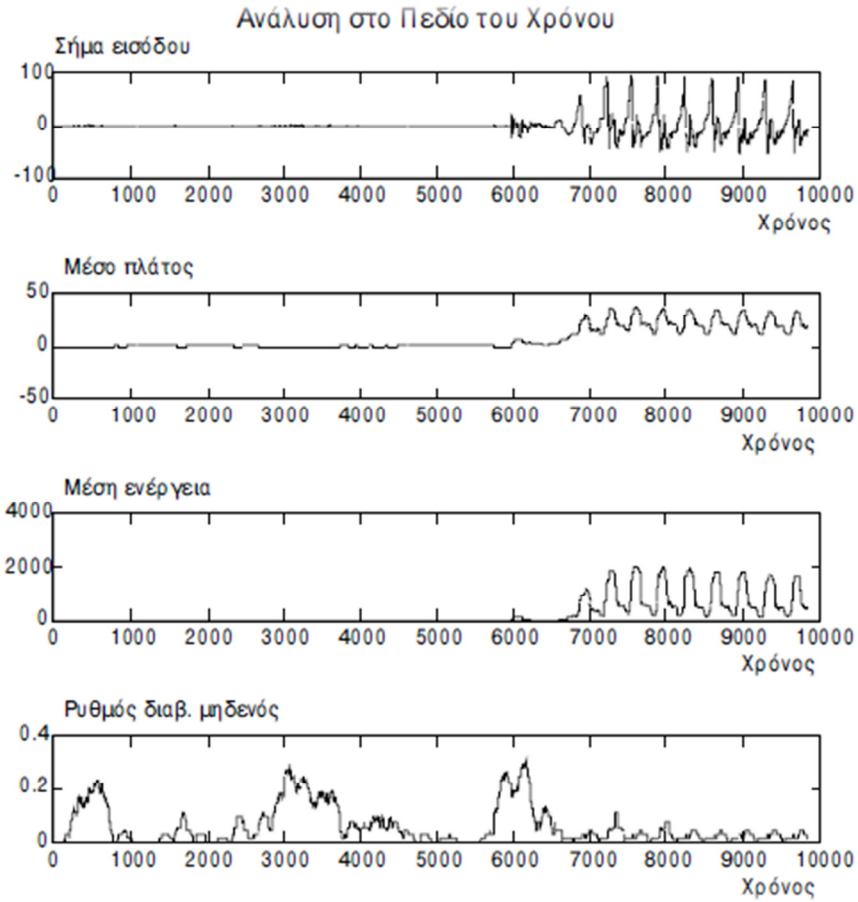
υπολογιστεί εύκολα με ένα συγκριτή προσήμου, ο οποίος θα συγκρίνει το τρέχον δείγμα για να διαπιστώσει αν αυτό έχει θετική ή αρνητική τιμή. Το αποτέλεσμα θα είναι 0 για δείγματα εισόδου πάνω από το μηδέν και 1 για δείγματα μικρότερα από το μηδέν.

Ο ρυθμός των εναλλαγών είναι:

$$Z(n) = \frac{1}{2N} \sum_{m=-\infty}^{+\infty} \left|sign\big(x(m)\big) - sign\big(x(m-1)\big)\right| w(n-m)$$

Το πρόβλημα της εύρεσης του ρυθμού εναλλαγών από το μηδέν είναι ότι επηρεάζεται από τις συνιστώσες ή από θόρυβο, λόγω του ότι η στάθμη του σήματος στα σύμφωνα είναι πολύ χαμηλή. Η λύση στο πρόβλημα αυτό είναι η χρήση ενός υψιπερατού φίλτρου με συχνότητα διέλευσης τα 70 Hz, ώστε να αποκοπεί και ο θόρυβος από την τροφοδοσία (50 Hz) που πιθανόν να υπάρχει.

Ακολουθεί ένα συνοπτικό σχήμα με ένα δείγμα και τα αποτελέσματα της κάθε μιας από τις προηγούμενες μεθόδους στο δείγμα αυτό.



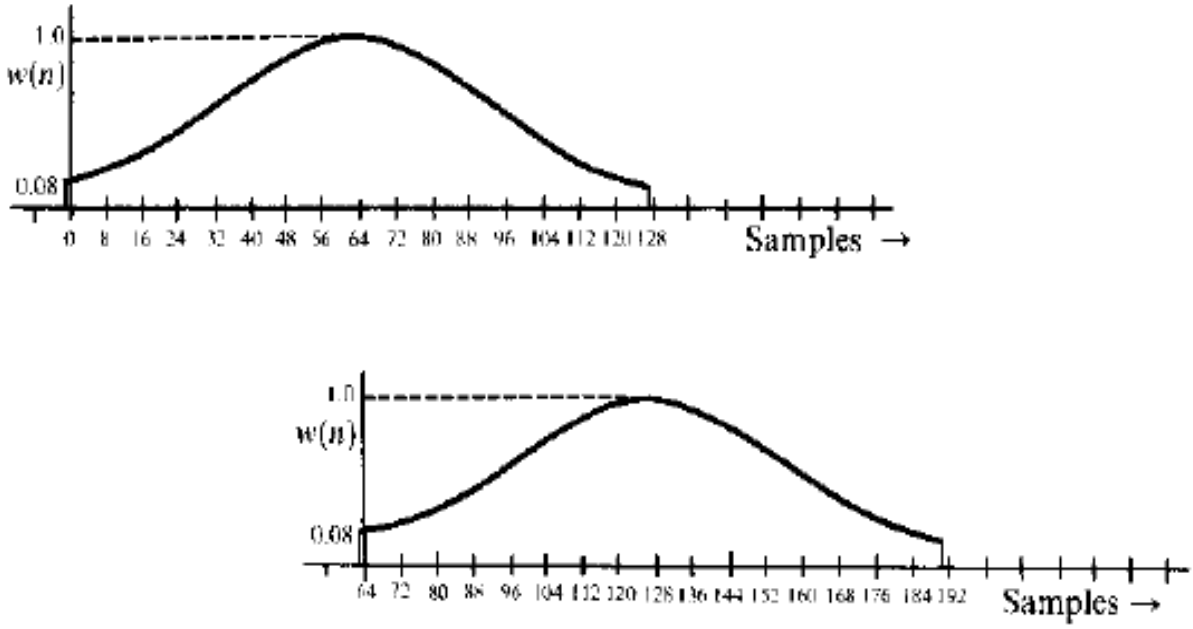
Εικόνα 1.13 Τμήμα μιας λέξης με τα χαρακτηριστικά του στο πεδίο του χρόνου

Επεξεργασία φθόγγων

Στο επίπεδο των φθόγγων γίνεται τμηματική ανάλυση σε αντίθεση με την πρωτογενή επεξεργασία όπου έχουμε επεξεργασία σε συνεχή ροή δειγμάτων. Αυτό σημαίνει ότι η επεξεργασία γίνεται σε μία ομάδα δεδομένων όπως για παράδειγμα το FFT (Fast Fourier Transform) όπου υπολογίζει τα συχνοτικά χαρακτηριστικά σε ομάδα δειγμάτων με μήκος, στην βέλτιστη περίπτωση, μία δύναμη του 2. Στην τμηματική ανάλυση μαζεύεται ένας πίνακας δειγμάτων, ο οποίος πολλαπλασιάζεται με μία συνάρτηση παραθύρου και κατόπιν ακολουθεί η κυρίως επεξεργασία. Ο πολλαπλασιασμός με το παράθυρο έχει σκοπό το φιλτράρισμα του σήματος εισόδου λόγω του ότι αυτό δεν είναι ένα συνεχές σήμα, αλλά ένα κομμάτι από αυτό, οπότε αν το θεωρήσουμε αυτόνομο, αλλοιώνονται τα συχνοτικά χαρακτηριστικά των δεδομένων. Για να είναι αυτή η αλλοίωση ελάχιστη χρησιμοποιούμε τα παράθυρα. Το τετραγωνικό παράθυρο δεν είναι και τόσο αποτελεσματικό αφού οι δευτερεύοντες λοβοί έχουν μεγάλη ισχύ σε σχέση με τον κύριο. Έτσι χρησιμοποιούμε άλλα παράθυρα όπως αυτά του Hanning, Hamming ή του Kaiser.

Βέβαια ένα παράθυρο πολλές φορές αν δεν πραγματοποιηθεί τη σωστή χρονική στιγμή μπορεί να απορρίψει κάποια σημαντικά χαρακτηριστικά από το σήμα εισόδου, όπως για παράδειγμα μπορεί να συμβεί με το γράμμα "τ". Το παράθυρο θα αποκόψει την απότομη εξαγωγή του αέρα οπότε και το φάσμα που θα προκύψει από την επεξεργασία, δεν θα περιέχει το φαινόμενο αυτό.

Αυτός είναι και ο λόγος που για να μην αποκόψουν πληροφορίες από το σήμα εισόδου, τα παράθυρα επεξεργάζονται ομάδες δειγμάτων που είναι αλληλοεπικαλυπτόμενες, έτσι ώστε κάθε δείγμα να βρίσκεται σε τουλάχιστον δύο ομάδες για επεξεργασία, όπως και στην εικόνα 1.14.



Εικόνα 1.14 Η μέθοδος τμηματικής ανάλυσης με αλληλεπικαλυπτόμενα παράθυρα

Οι επεξεργασίες τμηματικής ανάλυσης είναι

1) (Ταχύς) μετασχηματισμός Φουριέ (FFT)
2) Αυτό-συσχέτιση (Auto-covariance and correlation)
3) Πυκνότητα φάσματος (Power Spectral Density)
4) Cepstral analysis

Ο Μετασχηματισμός Φουριέ

Το φάσμα ενός περιοδικού σήματος που έχει υποστεί δειγματοληψία υπολογίζεται από το γνωστό διακριτό μετασχηματισμό του Φουριέ (DFT):

$$X(k) = \sum_{n=0}^{N-1} x(n)\, e^{-j2\pi nk/N}$$

Ο διακριτός μετασχηματισμός του Φουριέ δίνει μία ομάδα N τιμών που αντιστοιχούν στο φάσμα του σήματος. Επειδή η είσοδος που προέρχεται από φωνή έχει μόνο πραγματικό μέρος (πραγματική ακολουθία) το X(k) είναι συζυγής του X(N-k). Τα μέτρα των συζυγών είναι ίσα οπότε προκύπτει ότι το μέτρο του φάσματος από τα N/2 σημεία και άνω θα είναι συμμετρικό με τα πρώτα N/2 σημεία. Αυτό σημαίνει ότι αρκεί να απεικονίζουμε τα πρώτα N/2 σημεία του φάσματος αφού αυτά περιέχουν όλη την απαιτούμενη πληροφορία του σήματος για να το χαρακτηρίσουν.

Ο αντίστροφος μετασχηματισμός του Φουριέ είναι :

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k)\, e^{j2\pi nk/N}$$

Αν δεν χρησιμοποιηθούν όλα τα δείγματα ο αντίστροφος μετασχηματισμός δεν θα δώσει τις φανταστικές τιμές του αρχικού σήματος.

Η ανάλυση στο πεδίο των συχνοτήτων του μετασχηματισμού εξαρτάται από τη συχνότητα δειγματοληψίας και τον αριθμό των δειγμάτων :

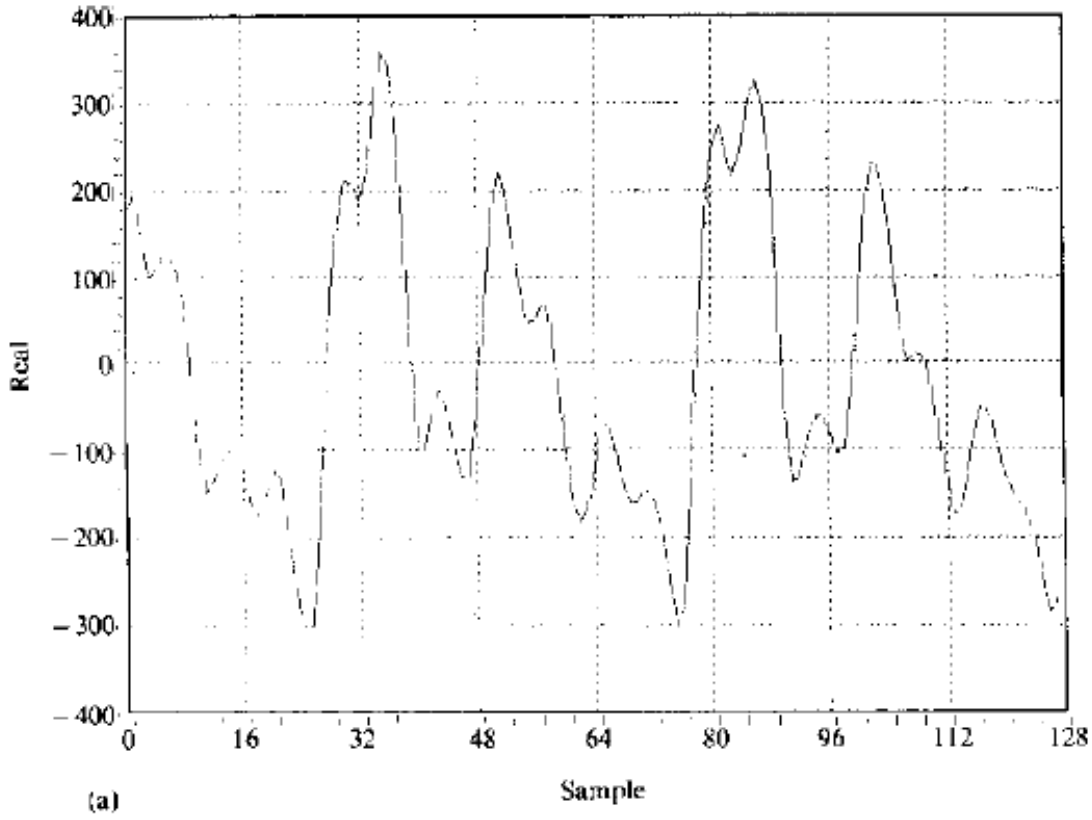
$$T = \frac{1}{f_s}$$

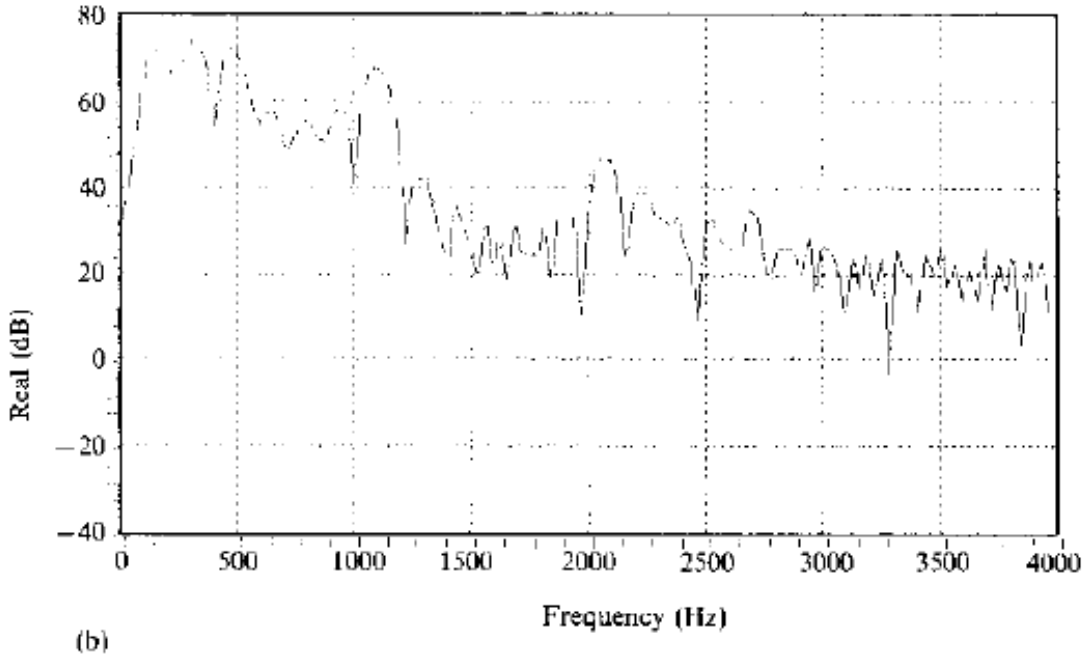
$$\Omega = \frac{1}{NT}\ (Hz) \text{ ή } \frac{2\pi}{NT}\ (rad/sec)$$

Στην εφαρμογή στην ομιλία υπάρχουν δυσκολίες στη χρήση λόγω του ότι η περιοδικότητα του σήματος μεταβάλλεται με τον χρόνο και μπορεί δύο συνεχόμενες ομάδες δειγμάτων φωνής να μην έχουν τις ίδιες συχνότητες, και το πλάτος της κάθε συχνότητας να είναι διαφορετικό.

Για να ξεπεραστούν τα παραπάνω προβλήματα χρησιμοποιούνται διάφορες τεχνικές εκ των οποίων οι πιο συνηθισμένες είναι η εύρεση του βέλτιστου μήκους της ομαδοποίησης, και η χρήση των παραθύρων για να αποφευχθούν τα φαινόμενα της διαρροής των συχνοτήτων. Για να ρυθμιστούν τα παράθυρα να έχουν ίδιο μήκος μπορούν να προστεθούν μηδενικές τιμές στο τέλος της ομάδας των δειγμάτων. Αυτό δεν αλλάζει τίποτα στο φάσμα σε μορφή, αν η ομάδα των δειγμάτων τελειώνει σε μηδέν, αλλά επειδή αυξάνεται ο αριθμός των δειγμάτων έχουμε μεγαλύτερη ανάλυση στο πεδίο των συχνοτήτων.

Αν παρατηρήσει κανείς το φάσμα ενός σήματος φωνής όπως το παρακάτω, μπορεί να εντοπίσει την βασική συχνότητα του σήματος όπως φαίνεται στην εικόνα 1.15.





Εικόνα 1.15 Τμήμα λέξης και η ανάλυσή του κατά Fourier

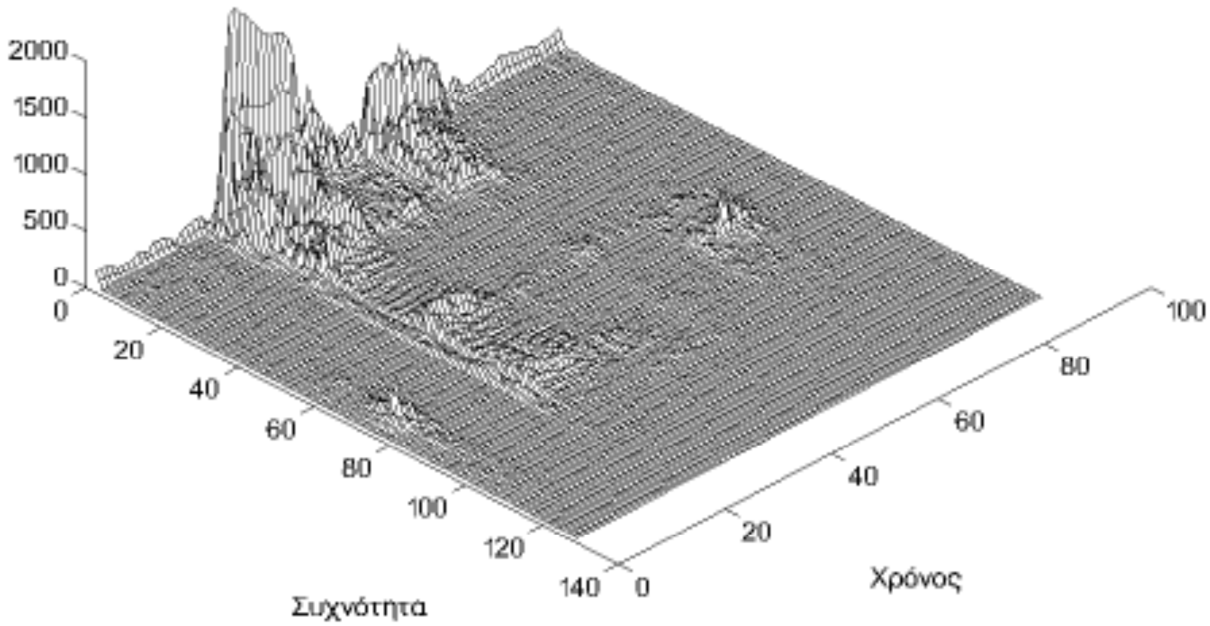
Στο δείγμα αυτό ο βασικός λοβός είναι στα 1050Hz, ο δευτερεύων στα 2030Hz, και λιγότερο καθαρά, οι επαναλήψεις ανά 160Hz σε όλο το φάσμα. Οι επαναλήψεις αυτές είναι η βασική συχνότητα της φωνής που είναι στα 160Hz. Επίσης διακρίνονται οι συχνότητες συντονισμού του φωνητικού συστήματος που φαίνονται ως αρμονικές της βασικής συχνότητας. Η πρώτη αρμονική έχει ευρύ φάσμα και είναι στα 320Hz. Η δεύτερη είναι στενότερη σε εύρος και είναι στα 1050Hz. Τέλος η τρίτη αρμονική βρίσκεται στα 2030Hz. Ανάμεσα στα 3 με 4 KHz η περιοδικότητα αλλάζει. Η αλλαγή αυτή οφείλεται στη συνάρτηση παραθύρου από τους δευτερεύοντες λοβούς του παραθύρου του Hamming.

Η φωνή έχει το χαρακτηριστικό ότι το πλάτος του σήματος στα σύμφωνα μειώνεται, με αποτέλεσμα να μειώνεται η ακρίβεια στις υψηλές συχνότητες (αφού τα σύμφωνα που τις περιέχουν έχουν χαμηλή ένταση). Για το λόγο αυτό χρησιμοποιούνται φίλτρα που ενισχύουν τις υψηλές συχνότητες ώστε να υπάρχει μία πιο επίπεδη απόκριση στο φάσμα. Τα φίλτρα αυτά καλούνται φίλτρα προέμφασης και δεν είναι τίποτε άλλο από ένα μηδενικό χαμηλών συχνοτήτων:

$$H_{preemph}(z) = 1 = az^{-1}, 0.95 < a < 0.98$$

Είναι γεγονός ότι ο διακριτός μετασχηματισμός του Φουριέ απαιτεί πολλές πράξεις και είναι απελπιστικά αργός σε εφαρμογές πραγματικού χρόνου. Για αυτό το λόγο χρησιμοποιείται στην πράξη μία παραλλαγή του, ο Ταχύς μετασχηματισμός Φουριέ, γνωστός και ως FFT. Ο μετασχηματισμός αυτός εξοικονομεί μεγάλο ποσό πράξεων (πολλαπλασιασμών και προσθέσεων) βασισμένος στη συμμετρία του DFT. Η χρήση της συμμετρίας έχει ως αποτέλεσμα ο FFT να απαιτεί την ομαδοποίηση των δειγμάτων σε δυνάμεις του 2 (δηλαδή ομάδες των 2,4,8,16 κοκ δειγμάτων), οπότε πρέπει να ληφθούν τα απαραίτητα μέτρα για την προσθήκη των μηδενικών και τη χρήση παραθύρων.

Μία εξελιγμένη χρήση του FFT είναι η παραγωγή των φασματογραμμάτων. Αυτά είναι τρισδιάστατα γραφήματα ενέργειας ανά συχνότητα και μονάδα χρόνου. Αυτά απεικονίζονται είτε τρισδιάστατα είτε ως δισδιάστατος πίνακας συχνότητας-χρόνου με την ισχύ σε κάθε συχνότητα να απεικονίζεται ως διαφορετικό ύψος ή ένταση του μελανιού ή με διαφορετικό χρώμα όπως φαίνεται στην εικόνα 1.16.



Εικόνα 1.16 Τρισδιάστατο φασματόγραμμα μιας λέξης

Τυπική απόκλιση, Αυτοσυσχέτιση (Autocovariance, Autocorrelation)

Σε περιοδικά σήματα φωνής, όπως τα φωνήεντα, η ανάλυση στο πεδίο του χρόνου δείχνει ότι υπάρχει κάποια σχέση μεταξύ του κάθε δείγματος και των γειτονικών του. Η τυπική απόκλιση για κάποια χρονική καθυστέρηση k δειγμάτων είναι ο μέσος όρος των γινομένων του κάθε δείγματος s(n) με το αντίστοιχό του k δείγματα πιο πέρα s(n+k). Η άθροιση αυτή πραγματοποιείται σε ένα μεγάλο εύρος δειγμάτων ώστε να καλυφθούν αρκετές περίοδοι από το σήμα της φωνής. Η τυπική αυτό-απόκλιση είναι:

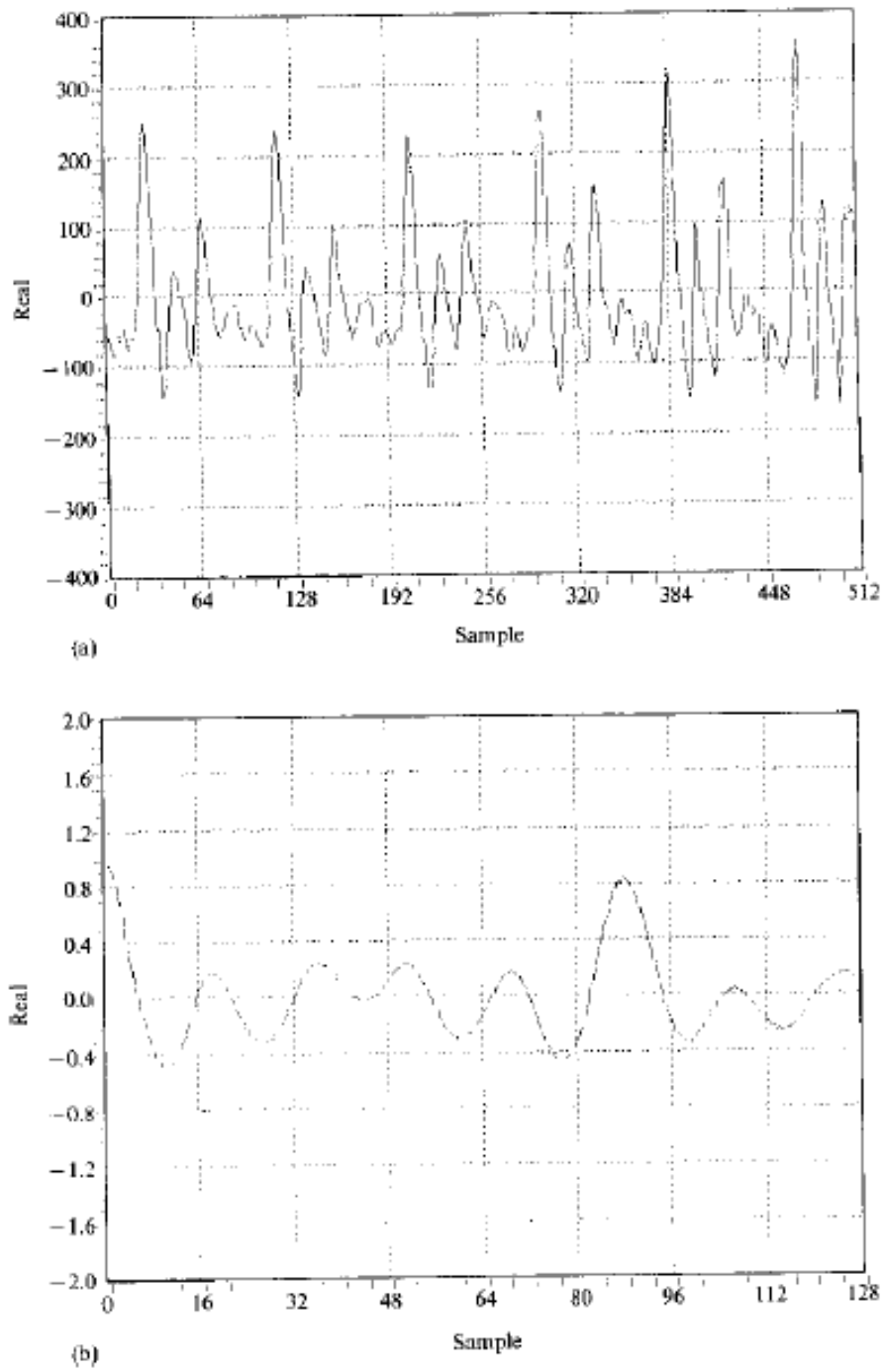
$$C(k) = \frac{1}{N}\sum_{n=0}^{N-1} s\,(n)s(n+k)$$

με μήκος ίσο με 0..N-1+k. Επειδή η τιμή του C(k) μπορεί να πάρει ποικίλες τιμές από πολύ μεγάλες μέχρι και πολύ μικρές για την σύγκριση μεταξύ λέξεων χρησιμοποιούμε κανονικοποίηση. Η κανονικοποίηση μας δίνει τη μέγιστη τιμή της όταν η καθυστέρηση k=0 οπότε έχουμε C(0)=1. Ο λόγος R(k):

$$R(k) = \frac{C(k)}{C(0)}$$

καλείται αυτοσυσχέτιση (autocorrelation). Αυτός είναι ο ορισμός της αυτοσυσχέτισης κατά τους Blackman and Tukey (1958) και μπορεί να βρεθεί με διάφορες παραλλαγές.

Μια κυματομορφή φωνής και η αυτοσυσχέτισή της φαίνονται στην εικόνα 1.17.



Εικόνα 1.17 Μια κυματομορφή και η αυτοσυσχέτισή της

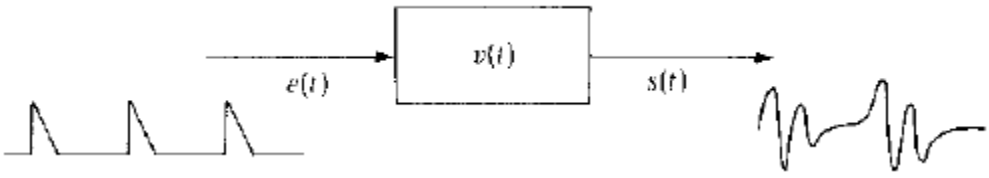
Το σήμα είναι ο φθόγγος /Q/. Η βασική συχνότητα είναι περίπου 90Hz με συχνότητα δειγματοληψίας τα 8KHz. Η αυτοσυσχέτιση δίνει τιμές κοντά στο 1 για πολύ μικρές καθυστερήσεις του k, πράγμα που σημαίνει ότι ένας αλγόριθμος πρόβλεψης μπορεί να προβλέψει την πορεία του σήματος για μικρές χρονικές περιόδους. Επίσης αφού παρουσιάζονται κορυφές με πλάτος κοντά στη μονάδα σε μακρινές αποστάσεις k_p που αντιστοιχούν στον τόνο της φωνής. Ο αλγόριθμος πρόβλεψης μπορεί να χρησιμοποιήσει και μεγάλες καθυστερήσεις για την πρόβλεψη του σήματος. Το πρόβλημα όμως είναι ότι αυτές οι καθυστερήσεις δεν είναι σταθερές και αλλάζουν με τον τόνο της φωνής οπότε πρέπει να υπάρχει προσαρμογή (adaptive system).

Ενεργειακή Πυκνότητα Φάσματος

Αν στη συνάρτηση της αυτοσυσχέτισης εκτελέσουμε μια ανάλυση κατά Φουριέ τότε παίρνουμε την πυκνότητα της ισχύς φάσματος. Η ισχύς του φάσματος είναι χρήσιμη διότι είναι μια πραγματική συνάρτηση της συχνότητας, και δείχνει πιο φανερά τα χαρακτηριστικά του σήματος σε σχέση με το πλάτος του φάσματος.

Cepstrum

Η φωνή αποτελείται από μια διέγερση που φιλτράρεται από τη στοματική κοιλότητα. Στο πεδίο του χρόνου αυτό σημαίνει συνέλιξη. Αν η είσοδος (διέγερση) είναι e(t), και η στοματική κοιλότητα έχει συνάρτηση μεταφοράς u(t) τότε το μοντέλο της ανθρώπινης ομιλίας φαίνεται στην παρακάτω εικόνα 1.18.



Εικόνα 1.18 Το μοντέλο παραγωγής της φωνής

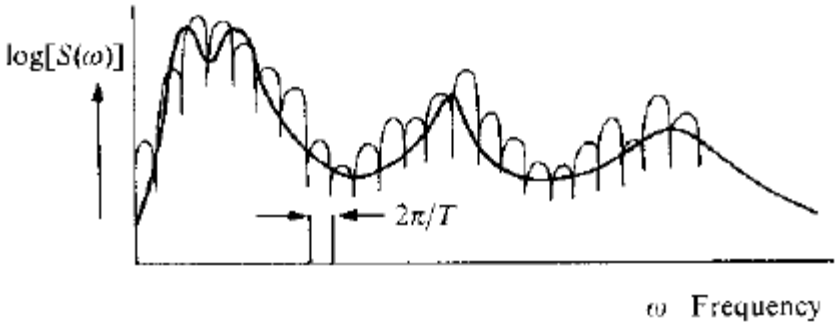
$$s(t) = e(t) * v(t)$$

Οπότε στο πεδίο των συχνοτήτων η συνέλιξη είναι πολλαπλασιασμός:

$$S(\omega) = E(\omega)V(\omega)$$

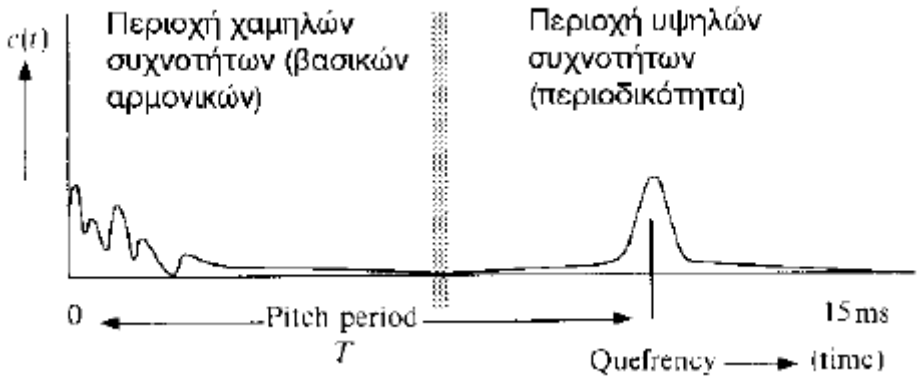
όπου S, E, U είναι οι μετασχηματισμοί κατά Φουριέ των s, e, v.

Τα χαρακτηριστικά της διέγερσης και της στοματικής κοιλότητας έχουν πολλαπλασιαστεί και είναι δύσκολο να διαχωριστούν. Αν όμως πάρουμε το λογάριθμο του S τότε ο πολλαπλασιασμός έχει μετατραπεί σε πρόσθεση δυο σημάτων όπως φαίνεται στην εικόνα 1.19.



Εικόνα 1.19. Ο λογάριθμος του φάσματος ενός τμήματος λέξης

Αν στο λογάριθμο αυτό εκτελέσουμε μια ανάλυση κατά Φουριέ , παίρνουμε το cepstrum του σήματος. Το cepstrum έχει διαστάσεις χρόνου. Στο διάγραμμα του cepstrum (με άξονα quefrency, "αντίστροφο" της συχνότητας) είναι φανεροί οι τόνοι που προέρχονται από τη διέγερση στα πρώτα ms ενώ ο τόνος της φωνής εμφανίζεται ως μια απότομη κορυφή δεξιότερα από τη μέση όπως και στην εικόνα 1.20.



Εικόνα 1.20 Το Cepstrum ενός τμήματος της λέξης

Αυτό σημαίνει ότι μπορούμε να φιλτράρουμε με μια συνάρτηση παραθύρου το cepstrum και να κάνουμε τον αντίστροφο μετασχηματισμό Φουριέ ώστε να έχουμε απομονώσει τη συνάρτηση μεταφοράς του στόματος. Από εκεί μπορούν να υπολογιστούν η συχνότητα, το πλάτος και το εύρος των βασικών αρμονικών.

ΚΕΦΑΛΑΙΟ 2

ΜΟΝΤΕΛΑ ΑΝΑΓΝΩΡΙΣΗΣ ΟΜΙΛΙΑΣ

Μέθοδοι Αναγνώρισης Φωνής

Η αναγνώριση φωνής είναι η διαδικασία που προσπαθεί να εξομοιώσει στον υπολογιστή την ακουστική αντιληπτική ικανότητα του ανθρώπου. Ήδη από την περιγραφή της διαδικασίας της αναγνώρισης, καταλαβαίνουμε ότι πρόκειται για κάτι εξαιρετικά πολύπλοκο, αφού καλείτε να προσεγγίσει μια από τις πολύπλοκες διαδικασίες της ανθρώπινης φύσης. Η διαδικασία της ακουστικής αντίληψης στον άνθρωπο, χρησιμοποιεί πολλά από τα μέρη του ανθρώπου, τόσο τα σωματικά όργανα (λάρυγγα, στόμα, πνεύμονες, νευρικό σύστημα) όσο και τα λογικά (μνήμη, αντίληψη, συσχετίσεις με άλλες αισθήσεις, κατανόηση). Αν σκεφτούμε ότι ακόμα δεν έχουμε κατανοήσει πλήρως τις διαδικασίες αυτές όπως συμβαίνουν στον άνθρωπο, γίνεται εύκολα αντιληπτό ότι δεν είναι δυνατόν, και ίσως δεν πρόκειται να γίνει, η πλήρης προσομοίωση της ικανότητας αυτής του ανθρώπου.

Λόγω του ενδιαφέροντος που έχει η αναγνώριση φωνής, έχει γίνει αντικείμενο μελέτης από πολλά χρόνια , ακόμη και πριν την διάδοση των ηλεκτρονικών υπολογιστών. Από την εποχή του Α΄ Παγκοσμίου πολέμου, οι επιστήμονες είχαν ξεκινήσει να καταγράφουν τα χαρακτηριστικά της ανθρώπινης ομιλίας, όπως αυτά περιγράφονται στο προηγούμενο κεφάλαιο, και να προσπαθούν να ερμηνεύσουν με το χέρι τα πρώτα φασματογράμματα, συλλεγμένα από πολλούς ομιλητές, σε λέξεις. Μετά, στην δεκαετία του ΄70, υπήρχαν αρκετοί επιστήμονες που ξεκίνησαν να επεκτείνουν την εργασία αυτή, έχοντας όμως σαν δυνατό εργαλείο τους τον ηλεκτρονικό υπολογιστή. Τα πρώτα επιτυχή αποτελέσματα ήρθαν γύρω στο 1978, όπου μπορούσαμε να έχουμε αναγνώριση κάποιας λέξης από τον υπολογιστή.

Βέβαια, οι στόχοι που έχουν τεθεί για την ιδανική περίπτωση αλγορίθμου αναγνώρισης, είναι να αναγνωρίζονται άπειρες λέξεις από οποιονδήποτε ομιλητή, σε συνεχή ροη λόγου και σε πραγματικό χρόνο. Αυτό άλλωστε, είχε φανεί και στα κινηματογραφικά έργα επιστημονικής φαντασίας, όπου το διαστημόπλοιο έπαιρνε τις εντολές για τον χειρισμό του από το στόμα του κυβερνήτη. Αν και υπάρχουν πολλά χρόνια μελέτης στο θέμα, σήμερα μπορούμε να έχουμε αλγορίθμους που λειτουργούν με αρκετά μεγάλο ποσοστό επιτυχίας (πάνω από 85%), με διάφορες μεθόδους. Μερικοί από τους πιο προχωρημένους αλγόριθμους λειτουργούν με αρκετά μεγάλη βάση δεδομένων (μερικές δεκάδες). Υπάρχουν φυσικά και ενδιάμεσοι τρόποι λειτουργίας, με βάση τους οποίους ταξινομούνται οι αλγόριθμοι αναγνώρισης, όπως

θα δούμε στην συνέχεια, μετά από μια σύντομη ματιά στα προβλήματα που παρουσιάζονται στην διαδικασία της αναγνώρισης.

Προβλήματα Στην Αναγνώριση

Στο προηγούμενο κεφάλαιο είδαμε ότι η δομή της ανθρώπινης ομιλίας είναι σχετικά απλή, με μερικές δεκάδες φθόγγους να συνθέτουν όλες τις λέξεις σε μια γλώσσα. Οι διάφοροι φθόγγοι παράγονται με κατάλληλες κινήσεις των πνευμόνων, του λάρυγγα, και την κατάλληλη άρθρωση του στόματος και την βοήθεια της ρινικής κοιλότητας. Όμως, το πρόβλημα είναι ότι υπάρχουν τεράστιες διαφορές στον τρόπο που ακούγονται αυτοί οι φθόγγοι, κυρίως λόγω των διαφορών που υπάρχουν στα φωνητικά όργανα ανάμεσα στους ανθρώπους.

Μεταξύ των διαφορών που παρατηρούνται είναι ότι η ανδρική φωνή είναι πιο "μπάσα", δηλαδή έχει πολλή ενέργεια στις χαμηλές συχνότητες, ενώ η γυναικεία έχει μεγάλη ενέργεια στις υψηλότερες συχνότητες. Επίσης, ανάλογα με την ταχύτητα με την οποία ομιλούμαι, διάφοροι φθόγγοι αλλάζουν σε διάρκεια, ακόμη και αν έχουν εκφωνηθεί από τον ίδιο ομιλητή, ή μπορεί και να παραλειφθούν εντελώς. Ακόμα, όταν μιλάμε, συνήθως έχουμε προετοιμάσει στο νου μας και τις επόμενες λέξεις που θα πούμε. Έτσι, όταν προφέρουμε ένα φθόγγο, ήδη τα φωνητικά όργανα έχουν αρχίσει να προετοιμάζονται να προφέρουν τον επόμενο. Αυτό κάνει δύσκολο τον διαχωρισμό των φθόγγων, αφού υπάρχει όχι μόνο επικάλυψη, αλλά και εξάρτηση των φθόγγων μεταξύ τους.

Αυτή η επικάλυψη, είναι ένα από τα πιο δύσκολα, ίσως το δυσκολότερο, από τα προβλήματα που πρέπει να αντιμετωπίσουμε στην αναγνώριση φωνής. Το πρόβλημα γίνεται πιο έντονο, όταν παρατηρήσουμε ότι στην ομιλία δεν υπάρχει σαφής διαχωρισμός ανάμεσα στις συλλαβές, αλλά ούτε και στις λέξεις. Ακόμη και η ακριβής εύρεση του σημείου που αρχίζει και του σημείου που τελειώνει χρονικά μια ομιλιθείσα λέξη, είναι μια διαδικασία δύσκολη και περιέχει συνήθως μεγάλο σφάλμα. Η δυσκολία αυξάνεται από τους θορύβους του περιβάλλοντος, γιατί σε αντίθεση με τον άνθρωπο, ο υπολογιστής δεν έχει την ικανότητα να συγκεντρώσει την προσοχή του σε κάποιον ομιλητή. Έτσι, ενώ ο άνθρωπος μπορεί να αντιληφθεί την ομιλία ακόμη και σε πολύ θορυβώδη περιβάλλοντα (πχ βιομηχανικούς χώρους, πάρτυ κλπ), απορρίπτοντας τους θορύβους του περιβάλλοντος και προσαρμοζόμενος σε αυτούς, ο υπολογιστής βρίσκεται χαμένος σε έναν ωκεανό παρασίτων με το S/N να πλησιάζει το μηδέν από τα αρνητικά.

Άλλο πρόβλημα που παρουσιάζεται είναι η διαφορά των φθόγγων που οφείλεται στην ομιλία διαφόρων διαλέκτων της ίδιας γλώσσας. Μερικοί φθόγγοι μπορεί να λείπουν ή να έχουν αντικατασταθεί από άλλους. Επίσης, διαφορές προκύπτουν και από την προσωδία, δηλαδή τον τρόπο με τον οποίο τονίζουμε τους φθόγγους μιας λέξης για να εκφράσουμε μηνύματα όπως απορία ή ερώτηση. Επιπλέον, κατά την εισαγωγή των σημάτων στον υπολογιστή, έχουμε διαφόρους τυχαίους θορύβους,

όπως από το ανοιγόκλεισμα των χειλιών, το πλατάγισμα της γλώσσας και τον θόρυβο από την κυκλοφορία του αέρα για την αναπνοή ή την εκφώνηση των φθόγγων.

Μια άλλη μεγάλη δυσκολία, είναι ότι πολλές λέξεις ακούγονται το ίδιο, αλλά δεν είναι ίδιες, όπως για παράδειγμα οι λέξεις “πολύ” και “πολλοί”. Εμείς, έχουμε συνηθίσει να τοποθετούμε την κάθε λέξη μέσα στην γλωσσική περίοδό της και με την βοήθεια του συντακτικού να καταλαβαίνουμε το σωστό νόημα. Μάλιστα πολλές φορές είμαστε σε θέση να προβλέψουμε μερικές φορές την λέξη που θα ακολουθήσει κάποια άλλη, έχοντας την γνώση του συντακτικού και της γραμματικής.

Είδη Αλγορίθμων

Για να υλοποιηθεί η αναγνώριση, πρέπει να γίνουν κάποιοι συμβιβασμοί ως προς το ιδανικό σύστημα. Έτσι, ανάλογα με τους περιορισμούς που δεχόμαστε στα διάφορα μέρη του συστήματος, δημιουργούνται διάφορες κατηγορίες αλγορίθμων. Μπορεί ο αλγόριθμος να λειτουργεί για έναν χρήστη, ή να καταλαβαίνει λίγες μόνο λέξεις ή να χρειάζεται να υπαγορεύουμε τις λέξεις αντί να τις ομιλούμαι κανονικά, ή ακόμα ο χρήστης να πρέπει να χρησιμοποιεί και μια ιδιότυπη γλώσσα με πολύ απλό συντακτικό και περιορισμένο λεξιλόγιο, που θα τα γνωρίζει ο υπολογιστής.

Έτσι λοιπόν, δημιουργείται ένας τρόπος με τον οποίο μπορούν να ταξινομηθούν οι διάφοροι αλγόριθμοι. Οι συνηθισμένες κατηγορίες και τα κριτήρια με τα οποία ταξινομούνται, είναι:

Με βάση τον τρόπο ομιλίας
- Σε ρυθμό φυσικής ομιλίας (συνεχούς)
- Σε ρυθμό υπαγόρευσης λέξεων

Με βάση το είδος της βάσης δεδομένων
- Για ένα χρήστη
- Για απεριόριστους χρήστες

Με βάση τον αλγόριθμο
- Με επεξεργασία φθόγγων
  - Με φωνητική ανάλυση
  - Με φασματική ανάλυση
- Με μοντελοποίηση
  - Hidden Markov Models
  - Χρήση LPC
- Με τεχνητή νοημοσύνη
  - Νευρωτικά δίκτυα
  - Ασαφής λογική

Το μοντέλο Linear Predictive Coding (LPC)

Η Γραμμική Προβλεπτική Κωδικοποίηση (LPC) είναι μία από τις πιο ισχυρές τεχνικές ανάλυσης και κωδικοποίησης ομιλίας καλής ποιότητας σε χαμηλό ρυθμό bit και παρέχει εξαιρετικά ακριβείς εκτιμήσεις των παραμέτρων ομιλίας. Η μέθοδος LPC μοντελοποιεί το ανθρώπινο φωνητικό σύστημα θεωρώντας ότι ένα σήμα ομιλίας παράγεται από ένα βόμβο στο τέλος ενός ηχητικού σωλήνα με την περιστασιακή προσθήκη συριστικών και εκρηκτικών ήχων. Αν και φαινομενικά αργό, το μοντέλο LPC αποτελεί μια καλή προσέγγιση της παραγωγής της ομιλίας. Η γλωττίδα παράγει το βόμβο, ο οποίος περιγράφεται από την ένταση και συχνότητα (pitch). Ο λαιμός και το στόμα σχηματίζουν το σωλήνα, που χαρακτηρίζεται από συντονισμούς (formants). Συριστικοί και εκρηκτικοί ήχοι δημιουργούνται από τη δράση της γλώσσας, των χειλιών και του λαιμού.

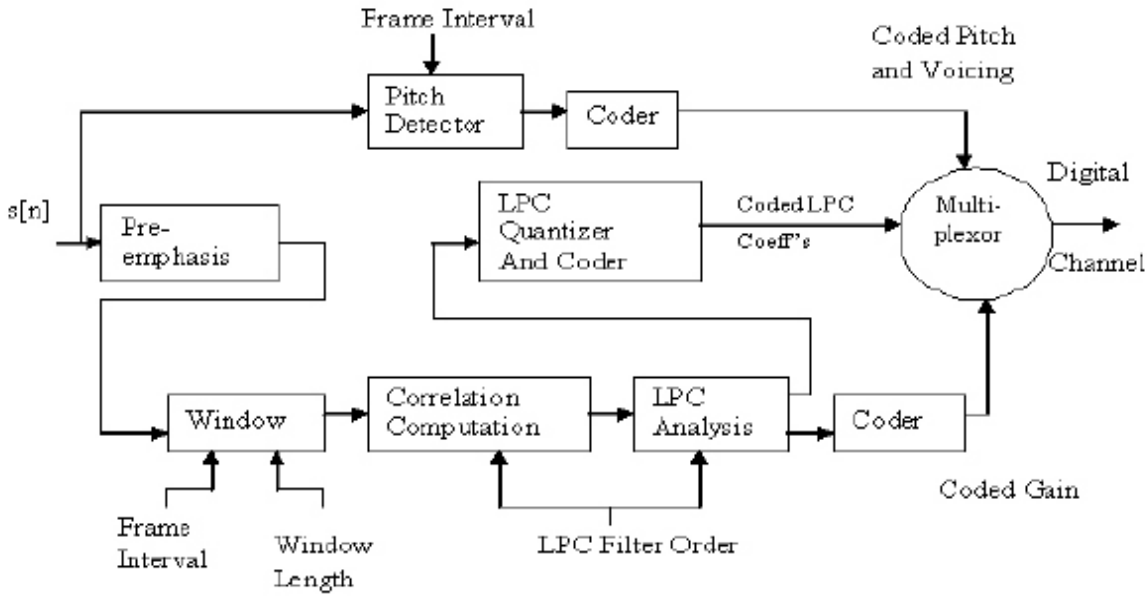
Διεγειρόμενος Θεμελιώδους Συχνότητας LPC (Pitch Excited LPC)

Όπως και όλοι οι άλλοι διεγερμένοι από τη θεμελιώδη συχνότητα vocoders, έτσι και ο διεγερμένος από τη θεμελιώδη συχνότητα LPC είναι ένας πλήρως παραμετροποιημένος κωδικοποιητής. Αυτό σημαίνει ότι η κωδικοποιημένη ομιλία χαρακτηρίζεται εξολοκλήρου από τις χρονικά μεταβαλλόμενες παραμέτρους ενός μοντέλου σύνθεσης ομιλίας. Αυτό το μοντέλο σύνθεσης έχει βασικά δύο τμήματα: το μοντέλο διέγερσης και το μοντέλο φωνητικού σωλήνα. Οι LPC τεχνικές χρησιμοποιούνται για να παραμετροποιηθεί σε αυτό το συνθέτη το μοντέλο του φωνητικού σωλήνα. Σε όλες τις τεχνικές κωδικοποίησης γραμμικής πρόβλεψης, η φωνητική περιοχή μοντελοποιείται ως ένα γραμμικά χρονικά μεταβαλλόμενο φίλτρο. Οι παράμετροι του γραμμικού φίλτρου παίρνονται μέσω μιας γραμμικής πρόγνωσης ανάλυσης του σήματος ομιλίας. Στους διεγειρόμενους από τη θεμελιώδη συχνότητα LPC's, το σήμα διέγερσης παραμετροποιείται πλήρως, και οι παράμετροι εξάγονται με τη χρήση ενός ανιχνευτή θεμελιώδους συχνότητας (pitch detector). Για άλλες κατηγορίες LPC's, η διέγερση αναπαριστάται και εξάγεται με διαφορετικούς τρόπους.

Τα Σχήματα 2.1 και 2.2 δείχνουν μπλοκ διαγράμματα ενός ολοκληρωμένου διεγειρόμενου από τη θεμελιώδη συχνότητα LPC αναλυτή (πομπού), και συνθέτη (δέκτη). Στον πομπό, οι παράμετροι του μοντέλου φωνητικού σωλήνα και οι παράμετροι του μοντέλου διέγερσης εξάγονται, κβαντίζονται, κωδικοποιούνται, πολυπλέκονται και μεταδίδονται. Στο δέκτη, οι κωδικοποιημένοι παράμετροι εξάγονται και χρησιμοποιούνται για τη σύνθεση της κωδικοποιημένης ομιλίας.

Ένας διεγειρόμενος από τη θεμελιώδη συχνότητα LPC πομπός κάνει δύο τύπων αναλύσεις: την ανάλυση διέγερσης (εύρεση θεμελιώδους συχνότητας) και την ανάλυση φωνητικού σωλήνα (LPC ανάλυση). Στο Σχήμα 2.1 ο ανιχνευτής θεμελιώδους συχνότητας βρίσκεται στο πάνω τμήμα του σχήματος και επενεργεί απευθείας στο σήμα εισόδου s[n] . Οι έξοδοι του ανιχνευτή θεμελιώδους συχνότητας περιέχουν μια φωνητική απόφαση (έμφωνο ή άφωνο) για κάθε πλαίσιο, και για τα έμφωνα πλαίσια μια περίοδο της θεμελιώδους συχνότητας. Αυτοί οι παράμετροι κωδικοποιούνται και πολυπλέκονται στην ροή δεδομένων εξόδου (output data stream).

Η LPC ανάλυση φαίνεται στο κάτω μισό του Σχήματος 2.1. στο τμήμα ανάλυσης, η ομιλία πρώτα περνάει από φίλτρο προέμφασης. Ο σκοπός αυτού του φίλτρου είναι να μειώσει το δυναμικό εύρος του φάσματος ομιλίας, το οποίο έχει ως αποτέλεσμα τη βελτιστοποίηση των αριθμητικών ιδιοτήτων των αλγορίθμων της LPC ανάλυσης. Μετά η ομιλία που έχει υποστεί προέμφαση παραθυροποιείται σε πλαίσια για ανάλυση. Ο τύπος παραθύρου, το μήκος παραθύρου, και το διάστημα μεταξύ δύο πλαισίων παραθύρου είναι βασικές παράμετροι ενός LPC κωδικοποιητή. Αφού έχει εφαρμοσθεί το παράθυρο, πραγματοποιείται μια ανάλυση συσχέτισης στα πεπερασμένου μήκους σήματα που έχουν προκύψει. Ο αριθμός των σημείων που χρησιμοποιούνται για την ανάλυσης συσχέτισης και ο σχετιζόμενος αριθμός των παραμέτρων που χρησιμοποιούνται για την LPC ανάλυση είναι οι κύριοι παράμετροι ελέγχου για τον συσχετιστή (correlator) και τον υποακολουθικό (subsequent) LPC αναλυτή. Τα αποτελέσματα της LPC ανάλυσης για κάθε πλαίσιο είναι η παράμετρος κέρδους και μια ομάδα παραμέτρων του LPC φίλτρου. Και οι δύο αυτές παράμετροι κβαντίζονται, κωδικοποιούνται, και πολυπλέκονται σε μια έξοδο ροής δεδομένων για εκπομπή ή αποθήκευση.

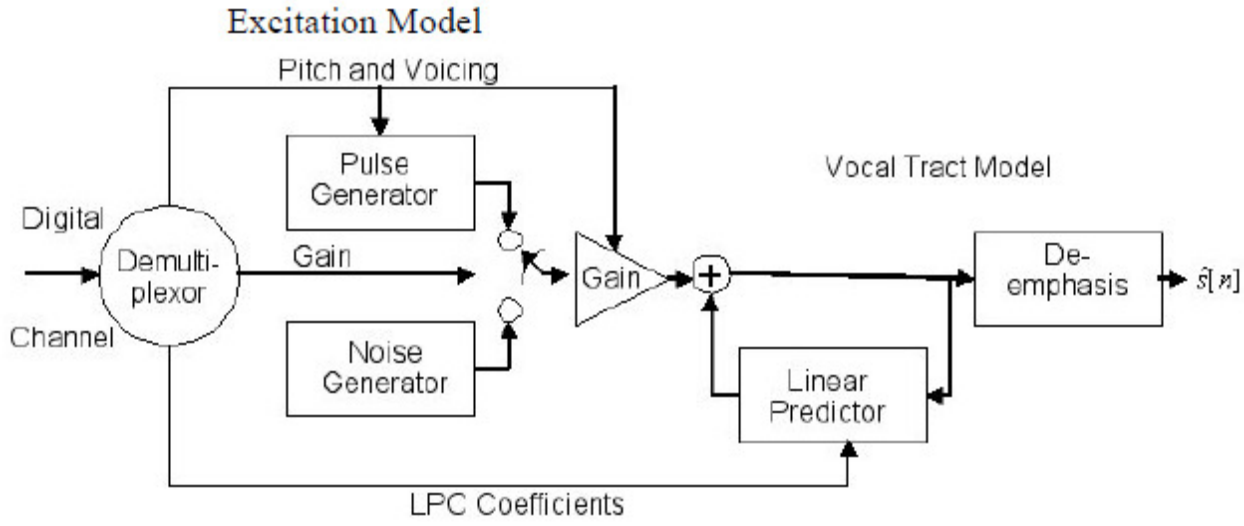


Σχήμα 2.1 Μπλοκ διάγραμμα ενός διεγερμένου από τη θεμελιώδη συχνότητα LPC πομπού

Το μπλοκ διάγραμμα ενός δέκτη γραμμικής πρόβλεψης διεγειρόμενου από τη θεμελιώδη συχνότητα vocoder φαίνεται στο Σχήμα 2.2. Ο βασικός συνθέτης ομιλίας αποτελείται από ένα σήμα διέγερσης που είναι μια είσοδος σε ένα χρονικά μεταβαλλόμενο φίλτρο φωνητικού σωλήνα. Η γεννήτρια διεγέρσεων περιλαμβάνει μια γεννήτρια παλμών, μια γεννήτρια θορύβου, έναν επιλογέα έμφωνων-άφωνων, και το κέρδος. Το φίλτρο φωνητικού σωλήνα δημιουργείται από ένα γραμμικό προβλεπτή που λειτουργεί σε ένα περιοδικά επαναλαμβανόμενο κύκλο. Το φίλτρο από-έμφαση

(de-emphasis filter) είναι το αντίστροφο φίλτρο για το φίλτρο προ-έμφασης που βρίσκεται στον πομπό.

Η λειτουργία του δέκτη μπορεί να συνοψισθεί ως εξής. Δεδομένα από το ψηφιακό κανάλι εισαγωγής αποπλέκονται στα τρία παρακάτω στοιχεία: θεμελιώδης συχνότητα και έμφωνα, κέρδος, και LPC συντελεστές. Τα δεδομένα θεμελιώδους συχνότητας χρησιμοποιούνται για τον έλεγχο του ρυθμού παλμών στη γεννήτρια παλμών ενώ τα έμφωνα δεδομένα χρησιμοποιούνται για τον έλεγχο της θέσης του διακόπτη έμφωνων. Τα δεδομένα κέρδους χρησιμοποιούνται για το έλεγχο του πλάτους του σήματος διέγερσης, και έτσι της έντασης της ομιλίας στην έξοδο. Οι LPC συντελεστές χρησιμοποιούνται για τον έλεγχο του φίλτρου φωνητικού σωλήνα. Ο ρόλος του φίλτρου από-έμφασης που ακολουθεί μετά το φίλτρο φωνητικού σωλήνα είναι να αναστρέψει τη φασματική προσαρμογή που είχε επιβληθεί στην ομιλία στον πομπό από το φίλτρο προ-έμφασης.



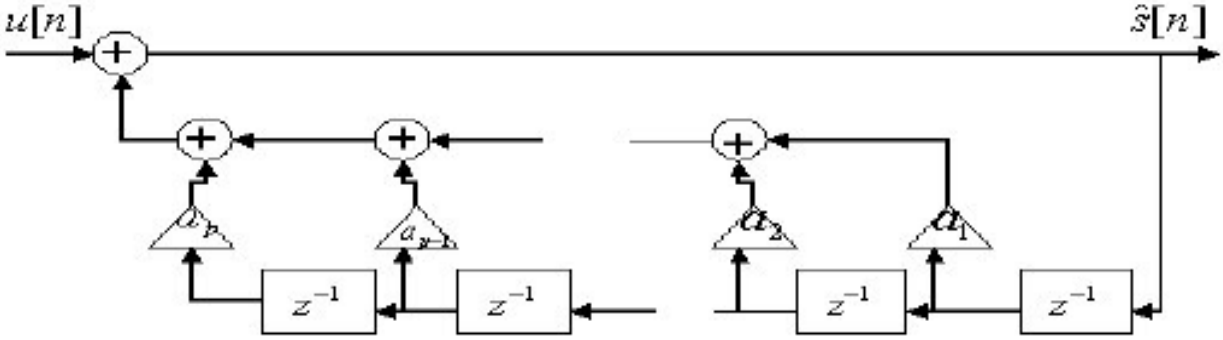
Σχήμα 2.2 Μπλοκ διάγραμμα ενός LPC δέκτη διεγερμένου από τη θεμελιώδη συχνότητα

Στο LPC μοντέλο, το φίλτρο σύνθεσης είναι μια αναπαράσταση του φαινομένου ακουστικού φιλτραρίσματος του φωνητικού σωλήνα. Το φίλτρο σύνθεσης συνήθως υλοποιείται ως ένα all-pole περιοδικά επαναλαμβανόμενο ψηφιακό φίλτρο του οποίου η είσοδο προσομοιάζει τη διέγερση στο φωνητικό σωλήνα και του οποίου η έξοδος είναι η συνθετική ομιλία.

Μοντέλο Φωνητικού Σωλήνα

Όπως φαίνεται και στο Σχήμα 2.2 ο συνθέτης ομιλίας που χρησιμοποιείται από τον LPC δέκτη μπορεί να διαιρεθεί σε δύο τμήματα: το μοντέλο διέγερσης και το μοντέλο φωνητικού σωλήνα. Το μοντέλο φωνητικού σωλήνα εμπεριέχει δύο στοιχεία: το φίλτρο φωνητικού σωλήνα και το φίλτρο από-έμφασης. Το φίλτρο φωνητικού σωλήνα μπορεί να υλοποιηθεί με διάφορες μορφές. Στην ποιο απλή υλοποίηση, ο φωνητικός σωλήνας μοντελοποιείται ως μια απευθείας μορφή ενός IIR φίλτρου όπως φαίνεται και στο Σχήμα 2.3. Σε όλες τις μορφές του, το φίλτρο φωνητικού σωλήνα χαρακτηρίζεται από P παραμέτρους, όπου P είναι συνήθως

μεταξύ 10-12 για ομιλία που έχει δειγματοληπτηθεί με 8000 δείγματα ανά δευτερόλεπτο.



Σχήμα 2.3 Απευθείας εφαρμογή του φίλτρου φωνητικού σωλήνα

Το κύριο έργο του πομπού όσο αναφορά το φίλτρο φωνητικού σωλήνα, είναι περιοδικά να αναλύει την ομιλία στην είσοδο (συνήθως 40-100 φορές ανά δευτερόλεπτο), για να υπολογίσει, να κβαντίσει, να κωδικοποιήσει και να εκπέμψει τις παραμέτρους του φωνητικού σωλήνα που είναι απαραίτητες για να υλοποιηθεί το φίλτρο φωνητικού σωλήνα στο δέκτη. Όπως φαίνεται και στο Σχήμα 2.1, αυτό επιτυγχάνεται σε τέσσερα βήματα: το φίλτρο προ-έμφασης, ο υπολογισμός της συσχέτισης, την LPC ανάλυση, και την LPC κβάντιση και κωδικοποίηση.

Υπολογισμός Συσχέτισης και η LPC Ανάλυση

Η LPC ανάλυση διεξάγεται πάνω σε πλαίσια δεδομένων. Η καρδία του LPC είναι ο γραμμικός προγνώστης. Στο γραμμικής πρόβλεψης μοντέλο, θεωρείται ότι το σήμα ομιλίας είναι μια αυτό-οπισθοδρομική (autoregressive) διαδικασία που μπορεί να αναπαρασταθεί ως

$$s(n) = \sum_{i=1}^{p} a_i\, s(n-1) + Gu(n), \qquad (1)$$

όπου $s(n)$ είναι η συνθετική ομιλία που παράγεται από το μοντέλο, $u(n)$ είναι το σήμα διέγερσης, $a_i\ i = 1, \dots, P$ είναι οι παράμετροι πρόγνωσης, και P είναι η τάξη του προγνώστη. Σε αυτή την έκφραση, G είναι η παράμετρος κέρδους που χρησιμοποιείται για να ταιριάζει την ενέργεια της συνθετικής ομιλίας με εκείνη του αρχικού σήματος ομιλίας. Στο πεδίο των z-μετασχηματισμών, $S(z)$ είναι η έξοδος του φίλτρου, $H(z)$ στο σήμα εισόδου, $U(z)$. Το LPC φίλτρο σύνθεσης δίνεται από τη σχέση

$$H(z) = \frac{1}{1 - A(z)}, \qquad (2)$$

όπου $A(z)$ το είναι το φίλτρο προγνώστη που δίνεται από

$$A(z) = \sum_{i=1}^{p} a_k \, z^{-k} \,, \qquad (3)$$

Με αυτούς τους όρους, το $S(z)$ μπορεί να γραφτεί ως εξής

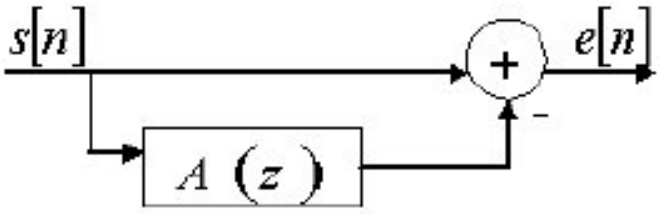
$$S(z) = H(z)U(z) = \frac{1}{\left(1 - A(z)\right)} = \frac{U(z)}{\left(1 - \sum_{k=1}^{p} a_k z^{-k}\right)}, \qquad (4)$$

Όπως φαίνεται και στο Σχήμα 2.2, το σήμα διέγερσης θεωρείται να είναι ένας παλμός εκπαίδευσης για την έμφωνη ομιλία και λευκός θόρυβος για την άφωνη ομιλία. Η περίοδος του παλμού είναι ίση με την περίοδο της θεμελιώδους συχνότητας του σήματος ομιλίας. Έτσι οι παράμετροι αυτού του μοντέλου σύνθεσης είμαι οι συντελεστές του προγνώστη (a_i , *s*), η περίοδος θεμελιώδους συχνότητας, η παράμετρος έμφωνου/άφωνου, και η παράμετρος κέρδους (*G*). Οι συντελεστές του προγνώστη είναι οι παράμετροι του φωνητικού σωλήνα, και οι υπόλοιπες είναι οι παράμετροι του σήματος διέγερσης.

Στην LPC ανάλυση ομιλίας, οι παράμετροι του μοντέλου διέγερσης και του μοντέλου φωνητικού σωλήνα προσεγγίζονται από το σήμα εισόδου ομιλίας. Όπως φαίνεται και από τη σχέση (4), οι μετασχηματισμοί της συνάρτησης μεταφοράς του φίλτρου φωνητικού σωλήνα και της διέγερσης πολλαπλασιάζονται μεταξύ τους στο πεδίο των *z*-μετασχηματισμών. Από την πλευρά του πεδίου συχνοτήτων, φαίνεται ότι το μοντέλο φωνητικού σωλήνα μεταφέρει την πληροφορία του φασματικού φακέλου, και το μοντέλο διέγερσης παρέχει την πληροφορία σχετικά με την φασματική λεπτομέρεια της ομιλίας.

Μοντέλο Χρονικά Μεταβαλλόμενου Φωνητικού Σωλήνα

Σε ένα LPC μοντέλο, ο φωνητικός σωλήνας αναπαριστάται από ένα all-pole φίλτρο $H(z)$. Επειδή η ομιλία είναι μια χρονικά μεταβαλλόμενη διεργασία, το $H(z)$ πρέπει να είναι ένα χρονικά μεταβαλλόμενο φίλτρο του οποίου οι συντελεστές μεταβάλλονται με το χρόνο. Επειδή ο φωνητικός σωλήνας κινείται σχετικά αργά, η ομιλία μπορεί να θεωρηθεί ότι είναι μια τυχαία διαδικασία της οποίας οι ιδιότητες μεταβάλλονται αργά. Αυτό οδηγεί στη βασική υπόθεση στατικότητας μικρού χρόνου που χρησιμοποιείται στην LPC ανάλυση. Αυτή η υπόθεση δηλώνει ότι το σήμα ομιλίας θεωρείται να είναι στατικό κατά τη διάρκεια ενός παραθύρου *L* δειγμάτων με την υπόθεση ότι το *L* είναι αρκετά μικρό. Αυτή η υπόθεση οδηγεί στη μοντελοποίηση της ομιλίας από διαδοχικά σταθερά φίλτρα $H(z)$'s, των οποίων οι συντελεστές παραμένουν σταθερές μέσα στο παράθυρο. Οι συντελεστές του $A(z)$, $a_i \; i = 1, \ldots, P$ παίρνονται μέσω ανάλυσης γραμμικής πρόγνωσης του σήματος ομιλίας.



Σχήμα 2.4 Μπλοκ διάγραμμα του αντίστροφου LPC φιλτραρίσματος

Υπάρχουν πολλοί τρόποι να δούμε την ανάλυση γραμμικής πρόγνωσης. Ένας από τους πιο διδακτικούς φαίνεται στο Σχήμα 2.4. Από αυτή την προοπτική, ο γραμμικός προγνώστης, $A(z)$, παράγει μια εκτίμηση του σήματος ομιλίας, $s(n)$, από το εισερχόμενο σήμα ομιλίας. Αυτή η εκτίμηση αφαιρείται από το αυθεντικό σήμα, δίνοντας ένα σήμα σφάλματος, $e(n)$,το οποίο ονομάζεται σήμα υπολοίπων πρόγνωσης. Αυτό το σήμα σφάλματος δημιουργείται από το αντίστροφο φίλτρο που δίνεται από

$$\frac{1}{H(z)} = 1 - A(z), \qquad (5)$$

Οι συντελεστές του προγνώστη υπολογίζονται από την ελαχιστοποίηση της ενέργειας των υπολοίπων πρόγνωσης, **E**, που δίνονται από τη σχέση

$$E = \sum_{n} e^2\,[n], \qquad (6)$$

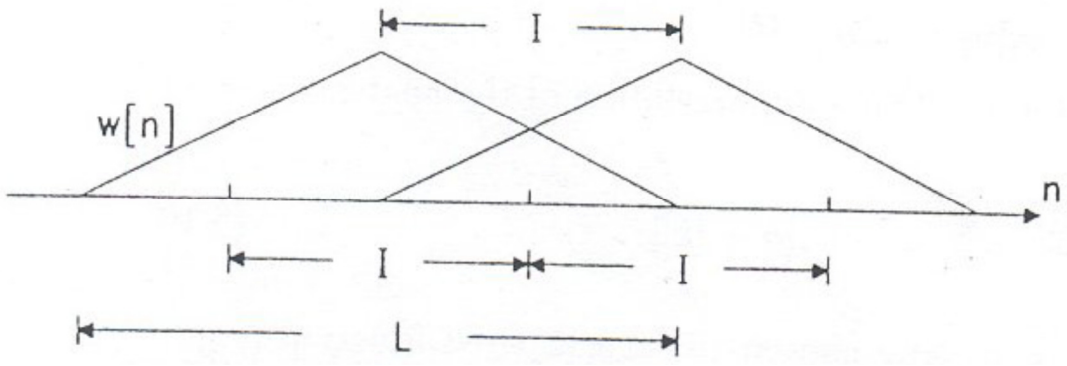
ως προς στους συντελεστές πρόγνωσης. Σε αυτή την έκφραση $e[n]$ είναι η έξοδος του αντίστροφου φίλτρου που δίνεται από

$$e[n] = s[n] - \sum_{i=1}^{p} a_i\, s[n-1], \qquad (7)$$

Υπάρχουν πολλοί μέθοδοι για να πάρουμε τους συντελεστές πρόγνωσης, οι βασικότεροι είναι η μέθοδος της αυτοσυσχέτισης και η μέθοδος covariance.

Μέθοδος Αυτοσυσχέτισης

Στη μέθοδο αυτοσυσχέτισης, ένα μετακινούμενο παράθυρο χρησιμοποιείται για να διαιρεθεί η ομιλία σε πλαίσια. Αυτή η διαδικασία φαίνεται στο Σχήμα 2.5. Για κάθε τοποθέτηση παραθύρου σε απόσταση 10 με 30 msec μεταξύ του, το σήμα ομιλίας παραθυροποιείται για να δημιουργηθεί ένα πλαίσιο ανάλυσης του σήματος.



Σχήμα 2.5 Τα κυλιόμενα παράθυρα εφαρμόζονται στο σήμα ομιλίας για την ανάλυση αυτοσυσχέτισης. Το μήκος παραθύρου L, είναι ανεξάρτητο από το διάστημα μεταξύ των πλαισίων, I.

Το σήμα που παράγεται είναι άπειρο σε έκταση, αλλά μηδέν οπουδήποτε εκτός του παραθύρου. Έτσι, είναι δυνατόν να υπολογιστεί η πραγματική συνάρτηση αυτοσυσχέτισης για ολόκληρο το σήμα. Το i^{th} πλαίσιο ανάλυσης δίνεται ως

$$s_i[n] = s[n]w_i[n], \qquad (8)$$

όπου $w_i[n]$ είναι το i^{th} πλαίσιο ανάλυσης. Το i^{th} πλαίσιο ανάλυσης συνήθως δίνεται από τη σχέση

$$w_i[n] = w[n - iI], \qquad (9)$$

όπου το I είναι το διάστημα ανάλυσης πλαισίου. Η αυτοσυσχέτιση του πλαισίου ανάλυσης καθορίζεται ως

$$R[|K|] = \sum_{-\infty}^{+\infty} s_i\,[n]s_i[n + |K|], \qquad (10)$$

Η συνάρτηση παραθύρου, $w[n]$, επιλέγεται να είναι μια συνάρτηση σταδιακής μείωσης (π.χ. ένα παράθυρο Hamming) μήκους L, όπου το L είναι το μέγεθος του παραθύρου ανάλυσης. Η ελαχιστοποίηση της μέσης εναπομένουσας ενέργειας στον πίνακα κανονικών εξισώσεων

$$Ra = r, \qquad (11)$$

όπου $a = \{a_1, \dots, a_p\}$ είναι το διάνυσμα των LPC συντελεστών, και **R** είναι ο πίνακας των συντελεστών αυτοσυσχέτισης και καθορίζεται ως

$$R[i, j] = R[|i - j|] = \sum_{n=-\infty}^{+\infty} s_i\,[n]s_i[n - j + i], \qquad (12)$$

και $r = \{R[1], \dots, R[P]\}$. Ο πίνακας **R** είναι ένας συμμετρικός Toeplitz πίνακας που μπορεί να λυθεί αποτελεσματικά με τη χρήση του αλγόριθμου Durbin. Ο αλγόριθμος αυτός είναι περιοδικά επαναλαμβανόμενος και χρησιμοποιεί τη δομή του

Toeplitz πίνακα **R** για να επιλύσει αποτελεσματικά τους LPC συντελεστές. Αυτός ο αλγόριθμος μπορεί να συνοψισθεί από το παρακάτω σετ εξισώσεων:

$$E^0 = R[0], \qquad (13)$$

$$k_i = \frac{\left[R[i] - \sum_j^{i-1} a_j^{i-1} R[i-j]\right]}{E^{i-j}}, \qquad (14)$$

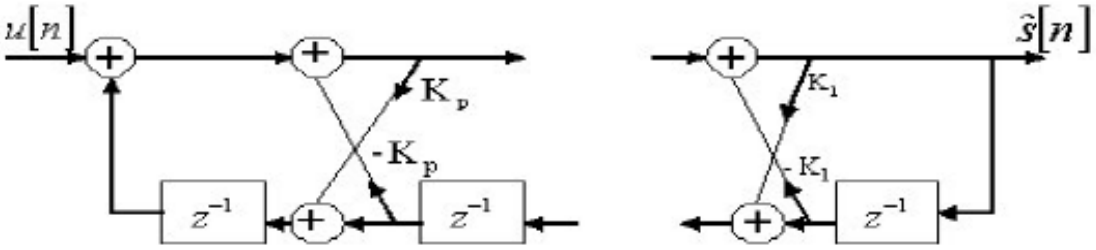
$$a_i^i = k_i, \qquad (15)$$

$$a_j^i = a_j^{i-1} + k_i a_{i-j}^{i-1}, \quad 1 \le j \le i-1 \ , (16)$$

$$E^i = (1 - k_i^2) E^{i-1} \qquad (17)$$

Οι εξισώσεις (13) και (14) λύνονται περιοδικά για $i = 1, \dots, P$. Οι συντελεστές k_i για $i = 1, \dots, P$ περιέχουν την ίδια πληροφορία με τους LPC συντελεστές, και ονομάζονται συντελεστές ανάκλασης (reflection coefficients) ή μερικοί συντελεστές συσχέτισης (γνωστοί και ως PARCORs). Το φίλτρο φωνητικού σωλήνα μπορεί να υλοποιηθεί απευθείας με τους PARCOR συντελεστές, όπως μπορούμε να δούμε και στο Σχήμα 2.6. Στην επίλυση για την τάξη του προγνώστη P, η περιοδικότητα παράγει όλους τους προγνώστες τάξης από 1 έως *P-1*. Η ποσότητα E^i είναι η ενέργεια του σφάλματος πρόγνωσης με προγνώστη τάξεως i. Καθώς E^i είναι μια θετική ποσότητα, η εξίσωση (17) μας δείχνει ότι όλοι οι PARCOR συντελεστές έχουν μέγεθος λιγότερο από ένα. Έτσι

$$-1 \le k_i < 1, \qquad (18)$$



Σχήμα 2.6 Δικτυωτή υλοποίηση του φίλτρου φωνητικού σωλήνα με τη χρήση των PARCORs

Επειδή το LPC φίλτρο φωνητικού σωλήνα είναι περιοδικό, η σταθερότητα είναι ένα πρόβλημα. Αλλά όπως φαίνεται η συνθήκη της εξίσωσης (18) είναι αρκετή για τη σταθερότητα του φίλτρου.

Μοντέλο Covariance

Στη μέθοδο covariance, το σήμα τη ομιλίας δεν παραθυροποιείται καθεαυτό, αλλά η ακολουθία του σφάλματος πρόβλεψης $e[n]$ από το Σχήμα 2.4 παραθυροποιείται και η ενέργεια του ελαχιστοποιείται. Έτσι η ποσότητα που καθορίζεται από

$$E = \sum_{-\infty}^{+\infty} e^2\,[n]w[n], \qquad (19)$$

ελαχιστοποιείται ως προς τους συντελεστές πρόγνωσης. Αυτή η ελαχιστοποίηση έχει ως αποτέλεσμα ένα πίνακα εξισώσεων της μορφής

$$\Phi \mathrm{a} = \phi, \qquad (20)$$

Όπου a είναι το διάνυσμα των συντελεστών πρόγνωσης, ο συμμετρικός πίνακας Φ καθορίζεται ως

$$\Phi = [i, j] = \sum_{n=-\infty}^{L-1} s[n-i]\,s[n-j], \qquad (21)$$

και $\phi = \{\Phi[1,0], \dots, \Phi[P,0]\}$. Καθώς το Φ δεν ένα Toeplitz πίνακας δεν μπορεί να επιλυθεί τόσο αποτελεσματικά σε σχέση με τις κανονικοποιημένες εξισώσεις της μεθόδου αυτοσυσχέτισης.

Τάξη Προγνώστη

Μια από τις αποφάσεις που πρέπει να παρθούν σε ένα LPC vocoder είναι η τάξη του LPC προγνώστη. Επειδή η ενέργεια που παραμένει μειώνεται με κάθε επανάληψη της Durbins recursion, η ενέργεια του σφάλματος πρόγνωσης μειώνεται καθώς ο αριθμός των πόλων του φίλτρου σύνθεσης, P, αυξάνεται. Καθώς ο αντικειμενικός σκοπός σε ένα vocoder είναι να εκπέμψει τους συντελεστές πρόγνωσης στο δέκτη, και λόγο του αριθμού των υπολογισμών, είναι σημαντικό να σταθεροποιηθούν και να περιορισθούν οι συντελεστές. Ένας τρόπος για να καθοριστεί το P το κατώφλι πέρα από το οποίο το σφάλμα δεν μειώνεται σημαντικά. Αν το κατώφλι είναι t_e , και αν

$$1 - \frac{E_{p+1}}{E_p} < t_e, \qquad (22)$$

τότε μια καλή επιλογή είναι $P = p$. Για ομιλία, δύο πόλοι (ένα πολικό ζεύγος) χρησιμοποιούνται για να μοντελοποιηθεί το κάθε formant. Το σήμα ομιλίας επίσης έχει φασματικά μηδενικά, αλλά επειδή αυτά έχουν ελάχιστες επιδράσεις, δεν μοντελοποιούνται στη συνάρτησης μεταφοράς του φωνητικού σωλήνα. Πρακτικά, για ομιλία 8 kHz, χρησιμοποιούνται τάξεις του προγνώστη σε ένα εύρος μεταξύ 10 και 16.

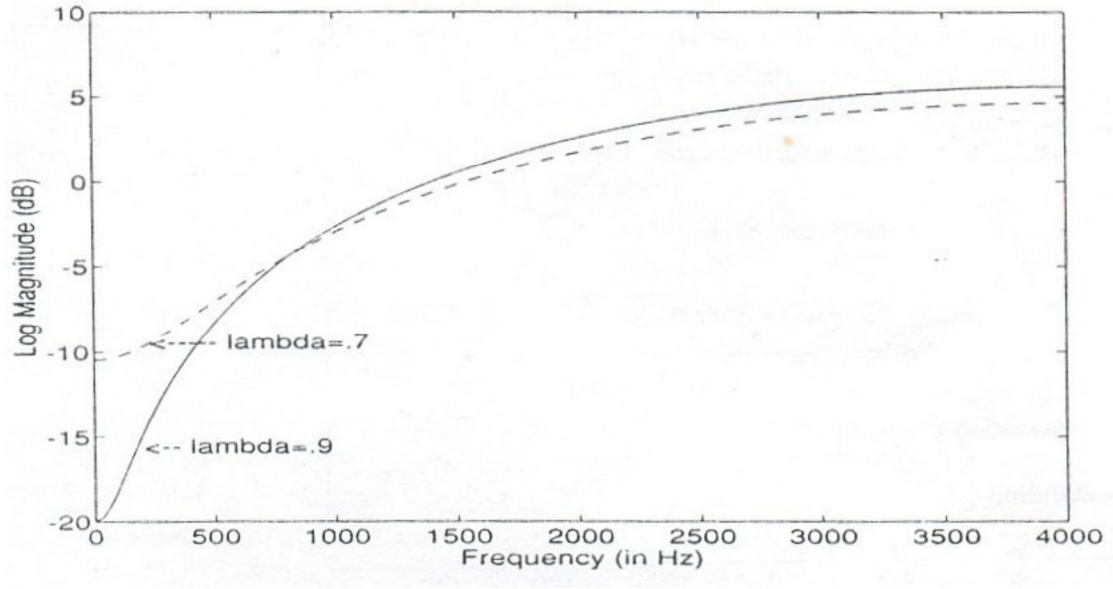
Προ-έμφαση

Το φάσμα έμφωνης ομιλίας συνήθως έχει μια πτώση κατά 6-db/octave, το οποίο έχει ως αποτέλεσμα υψηλά δυναμικό φασματικό εύρος. Αυτό έχει ως αποτέλεσμα το φάσμα ομιλίας να παρουσιάζει μια κλίση με τα υψηλότερα πλάτη να βρίσκονται στις χαμηλότερες συχνότητες (”το φάσμα έχει μια χαμηλοπερατή μορφή”). Αυτό το υψηλό δυναμικό εύρος συνήθως έχει ως αποτέλεσμα μια ανακριβή προσέγγιση των

υψηλότερων formants. Για να μειώσουμε αυτή την επίδραση, το αυθεντικό σήμα ομιλίας συχνά μπαίνει στη διαδικασία προ-έμφασης πριν από την LPC ανάλυση. Αυτό το σταθερό φίλτρο προ-έμφασης συνήθως έχει τη μορφή

$$V_{pre}(z) = 1 - \lambda z^{-1}, \qquad (23)$$

όπου $V(z)$ είναι αποτελεσματικό ήπιο υψιπερατό φίλτρο με ένα μηδενικό στο λ. Η σταθερά λ, ελέγχει το βαθμό προ-έμφασης. Το Σχήμα 2.7 δείχνει την απόκριση συχνότητας του φίλτρου προ-έμφασης για λ=0.7 και λ=0.9. Παρόλο που η βέλτιστη τιμή του λ μπορεί να υπολογιστεί στατιστικά, η τιμή διαφέρει για κάθε ομιλητή., και επιπλέον η ανάλυση δεν είναι ιδιαίτερα ευαίσθητη στην τιμή του λ.



Σχήμα 2.7 Απόκριση συχνότητας του φίλτρου προ-έμφασης για λ=0.7 και λ=0.9

Για να εξουδετερώσουμε την επίδραση της προ-έμφασης, στο δέκτη έχουμε ένα αντίστοιχο φίλτρο από-έμφασης της μορφής

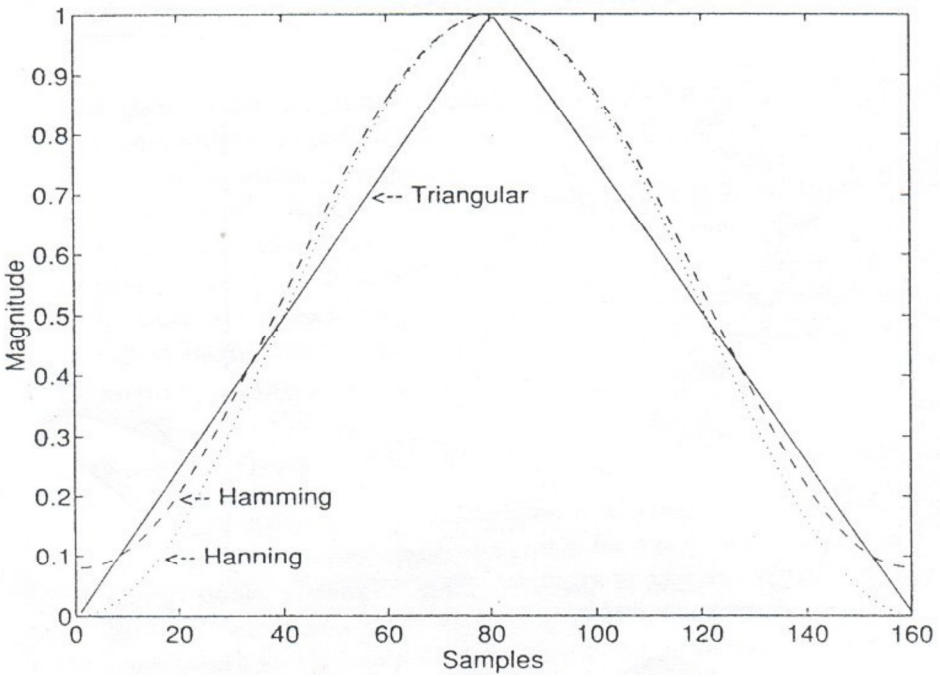
$$V_{de}(z) = \frac{1}{1 - \eta z^{-1}}, \qquad (24)$$

Παρόλο που το λ και το η επιλέγονται έτσι ώστε να εξουδετερώνει το ένα το άλλο, διαφορετικές τιμές του λ και η μπορούν να μας δώσουν καλύτερη ποιότητα ομιλίας.

Καθορισμός Παραθύρου

Ένα πολύ σημαντικό σετ παραμέτρων για τη γραμμικής πρόγνωσης ανάλυση είναι αυτές που απασχολούν τη λειτουργία του παραθύρου. Αυτές περιέχουν τον τύπο και το μέγεθος του παραθύρου που χρησιμοποιείται και το μέγεθος του διαστήματος του πλαισίου ανάλυσης. Ορισμένα τυπικά παράθυρα φαίνονται στο Σχήμα 2.8. Όταν

χρησιμοποιείται η μέθοδος αυτοσυσχέτισης, το παράθυρο εφαρμόζεται επανειλημμένως στο σήμα ομιλίας. Για να μειώσουμε τις επιδράσεις των άκρων του παραθύρου, χρησιμοποιούμε παράθυρα Hamming ή Hanning. Τέτοια ομαλά παράθυρα παράγουν καλύτερα αποτελέσματα από ορθογώνια παράθυρα ή παράθυρα με αιχμηρές άκρες. Το μέγεθος του παραθύρου, *L,* συνήθως επιλέγεται να καλύπτει μερικές περιόδους θεμελιώδους συχνότητας για έμφωνη ομιλία (20-40 msec). Αυτό είναι απαραίτητο για να μειώσουμε τις επιδράσεις του σήματος διέγερσης στην εκτίμηση των παραμέτρων του φίλτρου φωνητικού σωλήνα, και για να πάρουμε μια ποιο ακριβή εκτίμηση του φάσματος ομιλίας. Για τέτοια μεγέθη πλαισίων ανάλυσης, οι μέθοδοι αυτοσυσχέτισης και covariance παράγουν παρόμοια αποτελέσματα.



Σχήμα 2.8 Ορισμένα παράθυρα που χρησιμοποιούνται κατά την LPC ανάλυση. Το σχήμα δείχνει τα παράθυρα Hamming, Hanning και το τριγωνικό παράθυρο.

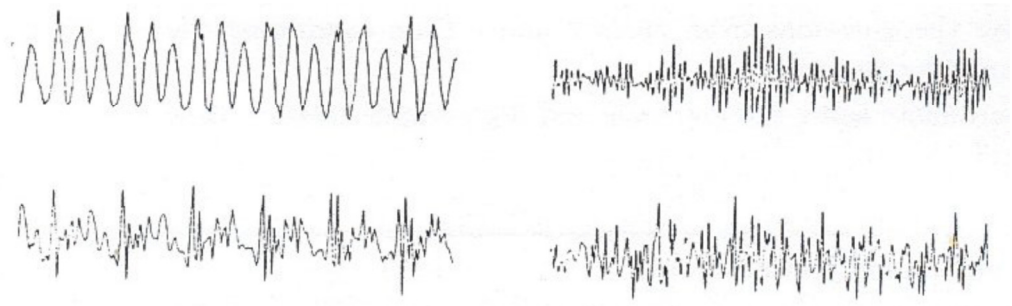
Το διάστημα πλαισίου ανάλυσης, *I,* καθορίζει τον αριθμό των δειγμάτων πάνω στα οποία θα χρησιμοποιηθούν οι LPC συντελεστές που προκύπτουν. Ο λόγος I/L αναπαριστά το ποσό της επικάλυψης μεταξύ δύο γειτονικών πλαισίων ανάλυσης (Σχήμα 2.5). Τυπικά χρησιμοποιείται μια διέγερση της τάξης του 50% (I=L/2). Ο Πίνακας 2.1 δείχνει όλες τις παραμέτρους της LPC ανάλυσης φωνητικού σωλήνα, το εύρος, και κάποιες τυπικές τιμές.

parameters	name	range	Typical values
predictor order	P	1-20	10
LPC window length	L	160-350	240
LPC frame size	I	40-160	120
Pre-emphasis factor	λ	0.7-0.95	0.8

Πίνακας 2.1 Οι παράμετροι της LPC ανάλυσης φωνητικού σωλήνα

Μοντέλο Διέγερσης

Υπάρχει ένας αριθμός δημοφιλών κωδικοποιητών γραμμικής πρόγνωσης που χρησιμοποιούνται σήμερα. Στο μεγαλύτερο μέρος τους, οι κωδικοποιητές αυτοί χρησιμοποιούν ένα μοντέλο γραμμικής πρόγνωσης φωνητικού σωλήνα, και οι περισσότεροι από αυτούς χρησιμοποιούν παρόμοιες τεχνικές LPC ανάλυσης.



Σχήμα 2.9 (αριστερά) Ένα τμήμα έμφωνης ομιλίας και από κάτω το ανταποκρινόμενο εναπομείναν σήμα. (δεξιά) Ένα τμήμα άφωνης ομιλίας και από κάτω το ανταποκρινόμενο εναπομείναν σήμα

Η βασική διαφορά μεταξύ αυτών των κωδικοποιητών είναι ο τρόπος με τον οποίο η είσοδος στο φίλτρο σύνθεσης $H(z)$ μοντελοποιείται και καθορίζεται. Για να καταλάβουμε τη φύση του σήματος διέγερσης σε ένα LPC περιβάλλον η εξίσωση (7) μπορεί να γραφτεί ως

$$s[n] = \sum_{i=1}^{p} a_i \, s[n-i] + e[n], \qquad (25)$$

Συγκρίνοντας τις εξισώσεις (1) και (25) είναι φανερό ότι αν $Gu[n] = e[n]$, τότε η έξοδος του $H(z)$ θα είναι ίση με την αυθεντική ομιλία. Έτσι, για να μπορέσει το LPC μοντέλο να παράγει μιας καλής ποιότητας συνθετική ομιλία, το $u[n]$ θα πρέπει να είναι μια καλή αναπαράσταση του εναπομείναντος σήματος $e[n]$. Το Σχήμα 2.9 μας δείχνει δύο τμήματα παραθύρων ενός σήματος ομιλίας και το ανταποκρινόμενο εναπομείναν σήμα για ένα προγνώστη $10^{ης}$ τάξης. Όπως μπορούμε να δούμε, το εναπομείναν σήμα για έμφωνη ομιλία είναι ένα ψευδό-περιοδικό σήμα, ενώ για την άφωνη ομιλία είναι ένα σήμα που ομοιάζει με θόρυβο. Στους διεγειρόμενους από τη θεμελιώδη συχνότητα LPC vocoders, το σήμα διέγερσης είναι πολύ απλό και αποτελείται είτε από περιοδικούς παλμούς είτε από λευκό θόρυβο. Έτσι λοιπόν, ένα απλό μοντέλο για το σήμα διέγερσης, $u[n]$, είναι να έχουμε μια εκπαιδευμένη διέγερση περιοδικών παλμών για την έμφωνη ομιλία και λευκό θόρυβο για την άφωνη ομιλία.

Για να παράγουμε ένα τέτοιο σήμα διέγερσης πρέπει να πάρουμε δύο παραμέτρους από το σήμα. Πρώτα, το αναλυόμενο πλαίσιο ομιλίας πρέπει να ταξινομηθεί ως έμφωνο ή άφωνο, και δεύτερον, για τα έμφωνα τμήματα πρέπει να καθοριστεί η περίοδος θεμελιώδους συχνότητας.

Ανίχνευση Θεμελιώδους Συχνότητας

Υπάρχουν πολλές προσεγγίσεις για να καθοριστεί η περίοδος της θεμελιώδους συχνότητας. Αυτές οι διαδικασίες μπορούν γενικά να διαιρεθούν στην προσέγγιση στο πεδίο του χρόνο και στην προσέγγιση στο πεδίο των συχνοτήτων. Στην προσέγγιση στο πεδίο του χρόνου, το σήμα ομιλίας επεξεργάζεται για να υπολογιστεί η περίοδος της θεμελιώδους συχνότητας. Στο πεδίο των συχνοτήτων, η φασματική πληροφορία και η αρμονική δομή του σήματος ομιλίας χρησιμοποιούνται για υπολογιστεί η περίοδος της θεμελιώδους συχνότητας.

Η πολυπλοκότητα και η ακρίβεια αυτών των προσεγγίσεων διαφέρουν σημαντικά ανάμεσα σε διαφορετικούς αλγορίθμους. Απλοί αλγόριθμοι, όπως ο κεντρικού ψαλιδίσματος και επιλογής κορυφής στην κυματομορφή της ομιλίας ή στην αντίστροφα φιλτραρισμένη ομιλία (υπόλοιπο πρόγνωσης) είναι παραδείγματα μεθόδων εύρεσης της περιόδου της θεμελιώδους συχνότητας στο πεδίο του χρόνου.

Υπολογισμός Κέρδους

Η παράμετρος κέρδους στο LPC μοντέλο χρησιμοποιείται για την παραγωγή ενός συνθετικού σήματος ομιλίας που έχει την ίδια ενέργεια με αυτή του αυθεντικού σήματος ομιλίας. Αυτό μπορεί να επιτευχθεί προσαρμόζοντας την ενέργεια της εξόδου του LPC φίλτρου για ένα παλμό (ή είσοδο λευκού θορύβου)στην ενέργεια του αυθεντικού σήματος ομιλίας. Αυτό έχει ως αποτέλεσμα την παρακάτω σχέση μεταξύ του κέρδους, και των συντελεστών αυτοσυσχέτισης του σήματος ομιλίας:

$$G = \left[R(0) - \sum_{k=1}^{p} a(k)R(k) \right]^{1/2}, \qquad (26)$$

Ο Πίνακας 2.2 μια λίστα LPC μοντέλων διέγερσης και παραμέτρων σύνθεσης.

parameters	name	range	typical values
predictor order	P	1-20	10
LPC frame size	I	40-160	120
de-emphasis factor	η	0.7-0.95	0.8

Πίνακας 2.2 Οι παράμετροι της LPC σύνθεσης

Κβαντισμός των Παραμέτρων του LPC Μοντέλου

Ένα σημαντικό στοιχείο όλων των LPC κωδικοποιητών είναι ο κβαντισμός και η κωδικοποίηση των παραμέτρων του μοντέλου σύνθεσης ομιλίας. Οι παράμετροι που μεταδίδονται διάστημα ανάλυσης είναι:

1. Συντελεστές προγνώστη $a_i : i = 1, \dots, P$
2. Περίοδος θεμελιώδους συχνότητας
3. Κέρδος
4. Παράμετροι έμφωνων

Η περίοδος θεμελιώδους συχνότητας, το κέρδος, και οι έμφωνες παράμετροι μπορούν να κβαντιστούν και να κωδικοποιηθούν με τη χρήση βαθμωτών κβαντιστών. Οι LPC συντελεστές προγνώστη να αναπαρασταθούν με διάφορες μορφές, κάποιες από τις οποίες είναι ποιο κατάλληλες για κβάντιση από άλλες. Η απευθείας κβάντιση των συντελεστών πρόγνωσης συνήθως αποφεύγεται λόγο του μεγάλου αριθμού bit που απαιτούνται για κάθε συντελεστή (απαιτούνται 8 με 10 bit). Τόσα πολλά bit απαιτούνται λόγω του ότι οι συντελεστές πρόγνωσης είναι πολύ ευαίσθητοι στα σφάλματα κβάντισης. Αυτό σημαίνει ότι μικρές διαφορές μπορεί να έχουν σημαντική επίδραση στο παραγόμενο φίλτρο σύνθεσης. Οι ισοδύναμες μορφές που είναι λιγότερο ευαίσθητες στη κβάντιση που έχουν προταθεί και χρησιμοποιούνται εμπεριέχουν:

1. Συντελεστές ανάκλασης $k_i\,'s$, (PARCORs)
2. Φασματικά ζεύγη γραμμής (LSPs), που ορίζονται να είναι ρίζες των πολυώνυμων $P(z)$ και $Q(z)$ που δίνονται από
$$P(z) = \big(1 - A(z)\big) + z^{-(P-1)}\big(1 - A(z^{-1})\big), \quad (27)$$
$$Q(z) = \big(1 - A(z)\big) - z^{-(P+1)}\big(1 - A(z^{-1})\big), \quad (28)$$
3. Τα πρώτα P δείγματα της κρουστικής απόκρισης των $H(z), h[n]$
4. Λόγοι λογαριθμικών περιοχών (LARs), που ορίζονται να είναι
$$LAR_i = \log\left(\frac{1 - k_i}{1 + k_i}\right), \qquad (29)$$
5. Συντελεστές αυτοσυσχέτισης, $R[i]'s$
6. Συντελεστές Cepstrum του $h[n]$, οι οποίοι μπορούν να παρθούν από την περιοδικά επαναλαμβανόμενη
$$h[n] = a_n + \sum_{k=1}^{n-1}\left(\frac{k}{n}\right)h[k]a_{n-k}, \qquad (30)$$

Υπολογισμός Φάσματος με τη Χρήση του LPC

Τεχνικές που βασίζονται στην ανάλυση γραμμικής πρόγνωσης έχουν εφαρμοστεί ευρέως για τον υπολογισμό του φάσματος για διάφορους τύπους σημάτων. Για σήματα ομιλίας, η απόκριση συχνότητας του φίλτρου σύνθεσης, $H(z)$, τείνει να ακολουθεί το φασματικό φάκελο του φάσματος ομιλίας. Αυτό μπορεί να φανεί εκφράζοντας το σφάλμα πρόγνωσης μέσου τετραγώνου στο πεδίο της συχνότητας. Στην πραγματικότητα, η γραμμική πρόγνωση μπορεί να διατυπωθεί στο πεδίο της συχνότητας, η οποία επίσης παράγει τις ίδιες κανονικοποιημένες εξισώσεις όπως φαίνεται στην εξίσωση (11).

Εφαρμόζοντας τον z-μετασχηματισμό στην εξίσωση (7) έχουμε

$$E(z) = \big(1 - A(z)\big)S(z), \qquad (31)$$

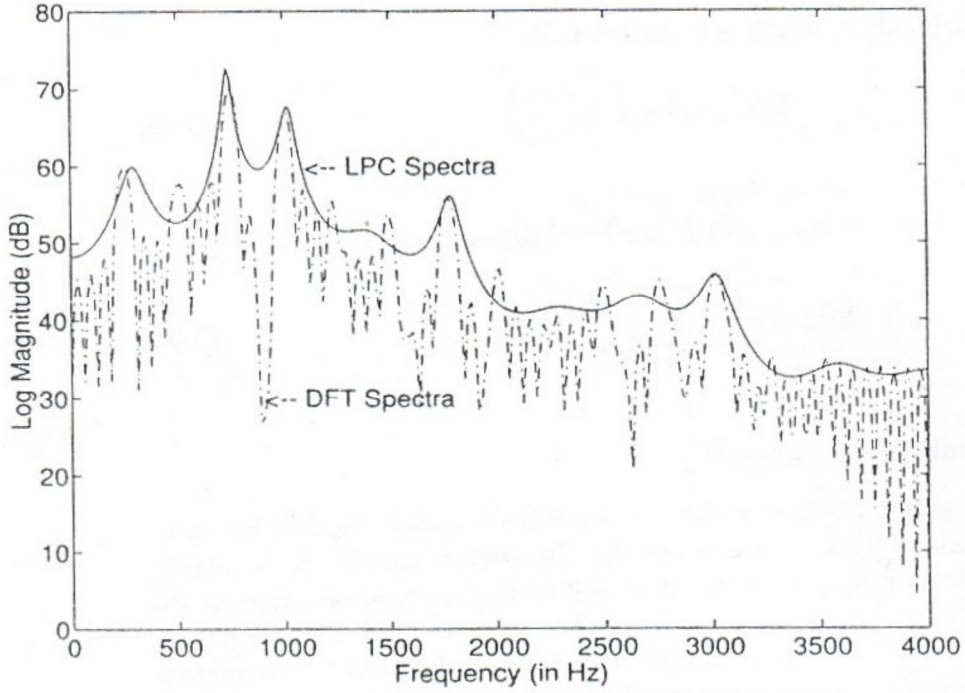
όπου $E(z)$ είναι ο z-μετασχηματισμός του υπολοίπου πρόγνωσης και $S(z)$ είναι ο z-μετασχηματισμός του σήματος ομιλίας. Χρησιμοποιώντας το θεώρημα του Parseval, το σφάλμα μέσου τετραγώνου μπορεί να εκφραστεί ως

$$E = \sum_n e^2\,[n] \int_{-\pi}^{\pi} |E\big(e^{j\omega}\big)\,|^2, \qquad (32)$$

Συνδυάζοντας τις εξισώσεις (31) και (32), το **E** μπορεί να εκφραστεί ως

$$E = \frac{G^2}{2\pi} \int_{-\pi}^{\pi} \frac{|S\big(e^{j\omega}\big)|^2}{|H(e^{j\omega})|^2} d\omega, \qquad (33)$$

Έτσι λοιπόν, ελαχιστοποιώντας το **E** είναι ισοδύναμο με το να ελαχιστοποιήσουμε το λόγο του ολοκληρώματος της φασματικής ενέργειας του σήματος ομιλίας ως προς τη φασματική ενέργεια της κρουστικής απόκρισης παλμού του φίλτρου, $H(z)$. Η εξίσωση (33) δείχνει τον τρόπο με τον οποίο το φάσμα του σήματος προσεγγίζεται από ένα φασματικό all-pole μοντέλο. Προφανώς, με την ελαχιστοποίηση του **E**, όπου ο λόγος των φασματικών ισχύων είναι μεγαλύτερος του 1 συνεισφέρουν περισσότερο στο συνολικό σφάλμα από τις περιοχές όπου ο λόγος είναι μικρότερος του 1. Έτσι το LPC φασματικό σφάλμα ευνοεί μια καλή αναπαράσταση των φασματικών κορυφών του σήματος. Αυτός είναι και ο λόγος που το $|H\big(e^{j\omega}\big)|^2$ συνήθως ακολουθεί το φασματικό φάκελο του $|S\big(e^{j\omega}\big)|^2$. Το Σχήμα 2.10 δείχνει ένα παράδειγμα του FFT φάσματος του σήματος, και έναν 20-pole LPC φασματικό υπολογισμό του σήματος. Στο Σχήμα 10 είναι δυνατό να δούμε τον τρόπο με τον οποίο το LPC φίλτρο συμπεριφέρεται ως φάκελος πάνω από την αρμονική δομή του σήματος διέγερσης.



Σχήμα 2.10 Το FFT φάσμα και το 20-pole LPC φάσμα ενός τμήματος του σήματος ομιλίας

ΚΕΦΑΛΑΙΟ 3

ΠΡΟΓΡΑΜΜΑΤΙΣΤΙΚΉ ΥΛΟΠΟΊΗΣΗ ΣΥΣΤΉΜΑΤΟΣ ΑΥΤΟΜΑΤΗΣ ΑΝΑΓΝΩΡΙΣΗΣ ΦΩΝΗΣ (MATLAB)

ΕΙΣΑΓΩΓΗ

Σε αυτό το κεφάλαιο θα γίνει η παρουσίαση ενός συστήματος αναγνώρισης φωνής σε περιβάλλον MATLAB καθώς και η επεξήγηση του κώδικα. Το προγραμματιστικό περιβάλλον του MATLAB προτιμάται λόγω της απλότητάς του σε διαχείριση μαθηματικών πράξεων και της ευκολίας που μας παρέχει στην εμφάνιση γραφικών παραστάσεων. Επίσης θα παρουσιαστεί μια πειραματική διαδικασία με διάφορα αρχεία φωνής που θα δοθούν σαν είσοδοι στο πρόγραμμα και έπειτα θα συγκριθούν τα αποτελέσματα.

3.1 ΚΩΔΙΚΑΣ ΥΛΟΠΟΙΗΣΗΣ ΠΡΟΓΡΑΜΜΑΤΟΣ

Η υλοποίηση του προγράμματος ξεκινά με την καταγραφή ενός αρχείου φωνής. Το αρχείο καταγράφεται με τη συνάρτηση **record_sound.m** ή **record_sound2.m** και με τη βοήθεια ενός μικροφώνου.

- Η **record_sound.m** καταγράφει μέσω μικροφώνου ένα αρχείο φωνής διάρκειας 4 sec.

```
function record_sound(filename)
disp('Test Sound recording (4 sec). Press any key to
continue');
pause;
Fs = 8000; % Sampling Freq (Hz)
Duration = 4; % Duration (sec)
y = wavrecord(Duration*Fs,Fs);%καταγραφή φωνής από μικρόφωνο
wavwrite(y,filename);
end
```

- Η **record_sound2.m** καταγράφει μέσω μικροφώνου ένα αρχείο φωνής με διάρκεια της επιλογής μας.

```
function record_sound(filename,Duration)
disp('Test Sound recording. Press any key to continue');
pause;
Fs = 8000; % Sampling Freq (Hz)
y = wavrecord(Duration*Fs,Fs);%καταγραφή φωνής από μικρόφωνο
wavwrite(y,filename);
end
```

Έχοντας καταγράψει το αρχείο φωνής, καλούμε την κύρια συνάρτηση του προγράμματος **lpc.m** η οποία κωδικοποιεί και αποκωδικοποιεί το αρχείο φωνής και εμφανίζει το αντίστοιχο ακουστικό αποτέλεσμα καθώς και την γραφική παράσταση αυτού.

- **lpc.m** κωδικοποίηση και αποκωδικοποίηση φωνής.

```
function lpc(filename,N,P)
%Κυρίως script

%Μεταβλητές εισόδου:
%       filename    - όνομα αρχείου ήχου
%       N           - μέγεθος δείγματος
%       P           - σύνολο συντελεστών

%Διάβασμα του αρχείου ήχου
%Στη μεταβλητή Fs αποθηκεύεται η συχνότητα δειγματοληψίας ενώ
στην
%x τα δείγματα
[x, fs] =wavread(filename);

%Θέλουμε το συνολικό χρόνο που διαρκεί το αρχείο ήχου, δηλαδή
%πόσα_δείγματα*κάθε_πότε_δειγματοληπτούμε (περίοδος
δειγματοληψίας=1/fs)
t=length(x)./fs;

%Καλούμε τη συνάρτηση κωδικοποίησης
%Θα μας επιστρέψει τους συντελεστές (θα τους χρειαστούμε στην
%αποκωδικοποίηση στη συνέχεια), την απόφαση αν έχει
%ομιλία και το κέρδος.
%Τα αποτελέσματα αφορούν το κάθε τμήμα ομιλίας.

%Το μήκος σε sec κάθε τμήματος υπολογίζεται ως fsize=N/fs
fsize=N/fs;
[lpc_coeffs, segment_pitch, voiced, gain] = lpc_encoder(x, fs,
P, fsize);
%Καλούμε τη συνάρτηση αποκωδικοποίησης και σύνθεσης
reconstructed = lpc_decoder (lpc_coeffs, segment_pitch, voiced,
gain);



%Αποτελέσματα
beep;
```

```
disp('Press any key to play the original');
pause;
soundsc(x, fs);

disp('Press any key to play the reconstructed LPC');
pause;
soundsc(reconstructed, fs);

figure;
subplot(2,1,1), plot(x); title(['Original = "', filename,
'"']);
subplot(2,1,2), plot(reconstructed); title(['Reconstructed "',
filename, '" with LPC']);
end
```

Στο κύριο πρόγραμμα παραπάνω καλείται η συνάρτηση **lpc_encoder.m** η οποία είναι υπεύθυνη για την κωδικοποίηση του αρχείου φωνής.

- **lpc_encoder.m** Κωδικοποίηση σήματος ομιλίας

```
function [lpc_coeffs, pitch, voiced, gain] = lpc_encoder(x, fs,
P, fsize)

%Μεταβλητές εισόδου:
%   x  - διάνυσμα δειγμάτων
%   fs - συχνότητα δειγματοληψίας
%   P  - σύνολο συντελεστών
%   fsize - διάρκεια τμήματος

%Μεταβλητές εξόδου:
%       aCoeff - συντελεστές για το κάθε τμήμα
%       pitch  - περίοδος διέγερσης κάθε τμήματος
%       v      - ομιλία ή όχι (0/1)
%       g      - κέρδος

%Υπολογισμός δειγμάτων ανά τμήμα
(ρυθμός_δειγματοληψίας*διάρκεια_τμήματος)
frame_length = round(fs .* fsize);
N = frame_length - 1;

%Συνάρτηση απόφασης για ύπαρξη ηχηρού τμήματος
%και υπολογισμού περιόδου διέγερσης
[voiced, pitch] = voiced_unvoiced (x, fs, fsize);

for b=1 : frame_length : (length(x) - frame_length)
    y1=x(b:b+N);
    y = filter([1 -.9378], 1, y1);  %pre-emphasis filtering

    %Εύρεση των συντελεστών με χρήση LEVINSON-DURBIN METHOD
    [autos,lags] = xcorr(y);
    autos = autos(find(lags==0):end);
    [a,g] = levinson(autos,P);

    lpc_coeffs(b: (b + P)) = a;
    e = y - filter([0 -lpc_coeffs(2:end)],1,y);
    %Υπολογισμός του gain για κάθε τμήμα
```

```
    gain(b) = segment_gain (e, voiced(b), pitch(b));
end
```

Στη συνάρτηση κωδικοποίησης εξετάζεται το κάθε τμήμα του ηχητικού μηνύματος για το αν είναι ηχηρό ή άηχο με τη βοήθεια της συνάρτησης **voiced_unvoiced.m**.

- **voiced_unvoiced.m** διάκριση ηχηρού-άηχου τμήματος

```
function [voiced, pitch] = voiced_unvoiced(x, fs, fsize)
%Μεταβλητές εισόδου:
%   x  - διάνυσμα δειγμάτων
%   fs - συχνότητα δειγματοληψίας
%   fsize - διάρκεια τμήματος

%Μεταβλητές εξόδου:
%       voiced - το τμήμα είναι ηχηρό ή όχι
%       pitch  - περίοδος διέγερσης κάθε τμήματος

frame_length = round(fs .* fsize);
N= frame_length - 1;

%Τμηματοποίηση σήματος εισόδου
for b=1 : frame_length : (length(x) - frame_length)
    y1=x(b:b+N);     %Προσωρινό τμήμα σήματος μήκους fsize
    y = filter([1 -.9378], 1, y1);  %pre-emphasis filter

    %Υπολογισμός ενέργειας σήματος
    msf(b:(b + N)) = segment_msf (y);
    %Υπολογισμός zero crossing
    zc(b:(b + N)) = segment_zc (y);
    %Υπολογισμός περιόδου διέγερσης
    pitch(b:(b + N)) = segment_pitch (y,fs);
end

thresh_msf = (( (sum(msf)./length(msf)) - min(msf)) .* (0.67) )
+ min(msf);
voiced_msf =  msf > thresh_msf;     %=1,0

thresh_zc = (( ( sum(zc)./length(zc) ) - min(zc) ) .*  (1.5) )
+ min(zc);
voiced_zc = zc < thresh_zc;

thresh_pitch = (( (sum(pitch)./length(pitch)) - min(pitch)) .*
(0.5) ) + min(pitch);
voiced_pitch =  pitch > thresh_pitch;

%Χρήση των παραπάνω κατωφλίων για αναγνώριση ηχηρού τμήματος
for b=1:(length(x) - frame_length)
    if voiced_msf(b) * voiced_pitch(b) * voiced_zc(b) == 1,
        voiced(b) = 1;
    else
        voiced(b) = 0;
    end
end
```

Για την διάκριση ηχηρού-άηχου τμήματος υπολογίσαμε την ενέργεια (**msf**) μέσω της συνάρτησης **segment_msf**, το zero crossing (**zc**) μέσω της συνάρτησης **segment_zc** και την περίοδο διέγερσης (**pitch**) με την συνάρτηση **segment_pitch** οι οποίες παρατίθενται παρακάτω.

- **segment_msf** υπολογισμός ενέργειας.

```
function res = segment_msf (y)
%Μεταβλητές εισόδου:
%   y  - διάνυσμα δειγμάτων τμήματος

%Μεταβλητές εξόδου:
%       msf - υπολογισμός ενέργειας σήματος

[B,A] = butter(9,.33,'low');
y1 = filter(B,A,y);

res=sum(abs(y1));
end
```

- **segment_zc** υπολογισμός zero crossing.

```
function res = segment_zc (y)

%Μεταβλητές εισόδου:
%   y  - διάνυσμα δειγμάτων τμήματος

%Μεταβλητές εξόδου:
%       ZC - zero crossing (μικρό=ηχηρό τμήμα ενώ μεγάλο=μη
ηχηρό)

res=0;

for n=1:length(y),
    if n+1<length(y)
        res=res + (1./2) .* abs(sign(y(n+1))-sign(y(n)));
    end
end
```

- **segment_pitch** υπολογισμός περιόδου διέγερσης.

```
function pitch_period = segment_pitch (y,fs)
%Μεταβλητές εισόδου:
%   y  - διάνυσμα δειγμάτων τμήματος
%   fs - συχνότητα δειγματοληψίας

%Μεταβλητές εξόδου:
%       ZC - zero crossing (μικρό=ηχηρό τμήμα ενώ μεγάλο=μη
ηχηρό)

%Καθορισμός του εύρους των δειγμάτων που αντιστοιχούν
%στη μέγιστη και ελάχιστη συχνότητα ομιλίας
period_min = round (fs .* 2e-3);
period_max = round (fs .* 20e-3);
```

```
%Μέθοδος αυτοσυσχέτισης
R=xcorr(y);

%Εντοπισμός θέσης μέγιστης τιμής αυτοσυσχέτισης (συχνότητας)
[~, R_mid]=max(R);

%Εξέταση του εύρους γύρω από τη μέγιστη τιμή
%Πρώτα όμως ελέγχουμε να μην ξεφύγουμε από τα όρια του
διανύσματος
if (R_mid + period_max>=length(R))
    period_max = length(R)-R_mid-1;
pitch_range = R ( R_mid + period_min : R_mid + period_max );
[~, R_mid] = max(pitch_range);

%Μας ενδιαφέρει η θέση της μέγιστης τιμής της αυτοσυσχέτισης
pitch_period = R_mid + period_min;
end
```

Τέλος, στη συνάρτηση κωδικοποίησης υπολογίζουμε και το κέρδος (**gain**) με τη συνάρτηση **segment_gain** που ακολουθεί.

- **segment_gain** υπολογισμός κέρδους.

```
function [gain] = segment_gain(e, voiced_b, pitch)

%Μεταβλητές εισόδου:
%   e              - σφάλμα για το τρέχον τμήμα
%   voiced_b       - το τμήμα είναι ηχηρό (0/1)
%   pitch   - περίοδος διέγερσης για το ηχηρό τμήμα

%Μεταβλητές εξόδου:
%       gain_b - gain για το τρέχον τμήμα

if voiced_b == 0
    power_b = mean(e.^2);
    gain = sqrt( power_b );
else
    denom = ( floor( length(e)./pitch ) .* pitch );
    power_b = sum( e (1:denom) .^2 ) ./ denom;
    gain = sqrt( pitch .* power_b );
end

end
```

Στη συνέχεια του κύριου προγράμματος καλείται η συνάρτηση **lpc_decoder.m** η οποία πραγματοποιεί την αποκωδικοποίηση του αρχείου ομιλίας.

- **lpc_decoder.m** Αποκωδικοποίηση σήματος ομιλίας.

```
%Αποκωδικοίηση-Σύνθεση
function synth_speech = lpc_decoder (lpc_coeffs, segment_pitch,
voiced, gain)
%Μεταβλητές εισόδου:
%   aCoeff      - διάνυσμα συντελεστών για όλα τα τμήματα
```

```
%   pitch_plot  - περίοδος διέγερσης για κάθε τμήμα
%   voiced      - ηχηρό τμήμα
%   gain        - gain κάθε τμήματος

%Μεταβλητές εξόδου:
%       synth_speech - σήμα μετά την ανάκτηση

%re-calculating frame_length for this decoder,
frame_length=1;
for i=2:length(gain)
    if gain(i) == 0,
        frame_length = frame_length + 1;
    else break;
    end
end

for b=1 : frame_length : (length(gain))
    %Ανάλογα με το αν το τμήμα έχει ανιχνευθεί ως ηχηρό ή όχι
    %καλούμε τις αντίστοιχες συναρτήσεις
    if voiced(b) == 1
        syn_y1 = voiced_synthesis (lpc_coeffs, gain(b),
frame_length, segment_pitch(b),b);
    else
        syn_y1 = unvoiced_synthesis (lpc_coeffs, gain(b),
frame_length, b);
    end

    synth_speech(b:b+frame_length-1) = syn_y1;
end
```

Για τη σύνθεση του σήματος ομιλίας στην παραπάνω συνάρτηση χρησιμοποιήθηκαν δυο συναρτήσεις. Η **voiced_synthesis** για τη σύνθεση των ηχηρών τμημάτων και η **unvoiced_synthesis** για τα άηχα τμήματα.

- **voiced_synthesis** Σύνθεση ηχηρού τμήματος

```
function syn_y1 = voiced_synthesis (lpc_coeffs, gain,
frame_length, pitch_plot_b,b)

%Μεταβλητές εισόδου:
%   aCoeff          - διάνυσμα συντελεστών  τμήματος
%   gain            - gain τμήματος
%   frame_length    - μήκος τμήματος
%   pitch_plot_b    - περίοδος διέγερσης τμήματος

%Μεταβλητές εξόδου:
%   syn_y1 - τμήμα σήματος μετά την ανάκτηση

%Δημιουργία ακολουθίας Kronecker
%η οποία θα χρησιμοποιηθεί ως σήμα εισόδου στον αποκωδικοποιητή
for f=1:frame_length
    if f./pitch_plot_b == floor(f./pitch_plot_b)
        ptrain(f) = 1;
    else ptrain (f) = 0;
    end
end
```

```
%Πέρασμα της ακολοθίας από ένα "αντίστροφο φίλτρο"
%για την ανάκτηση του ηχηρού τμήματος
syn_y2 = filter(1, [1 lpc_coeffs((b+1):(b+1+9))], ptrain);
syn_y1 = syn_y2 .* gain;
end
```

- **unvoiced_synthesis** Σύνθεση άηχου τμήματος

```
function syn_y1 = unvoiced_synthesis (lpc_coeffs, gain,
frame_length, b)

%Μεταβλητές εισόδου:
%   aCoeff          - διάνυσμα συντελεστών  τμήματος
%   gain            - gain τμήματος
%   frame_length    - μήκος τμήματος

%Μεταβλητές εξόδου:
%   syn_y1 - τμήμα σήματος μετά την ανάκτηση

%Δημιουργία θορύβου ως σήμα εισόδου για τον αποκωδικοποιητή
wn = randn(1, frame_length);

%Πέρασμα της ακολοθίας από ένα "αντίστροφο φίλτρο" με τους
συντελεστές που
%παρήγαγε ο κωδικοποιητής
syn_y2 = filter(1, [1 lpc_coeffs((b+1):length(b))], wn);
syn_y1 = syn_y2 .* gain;
end
```

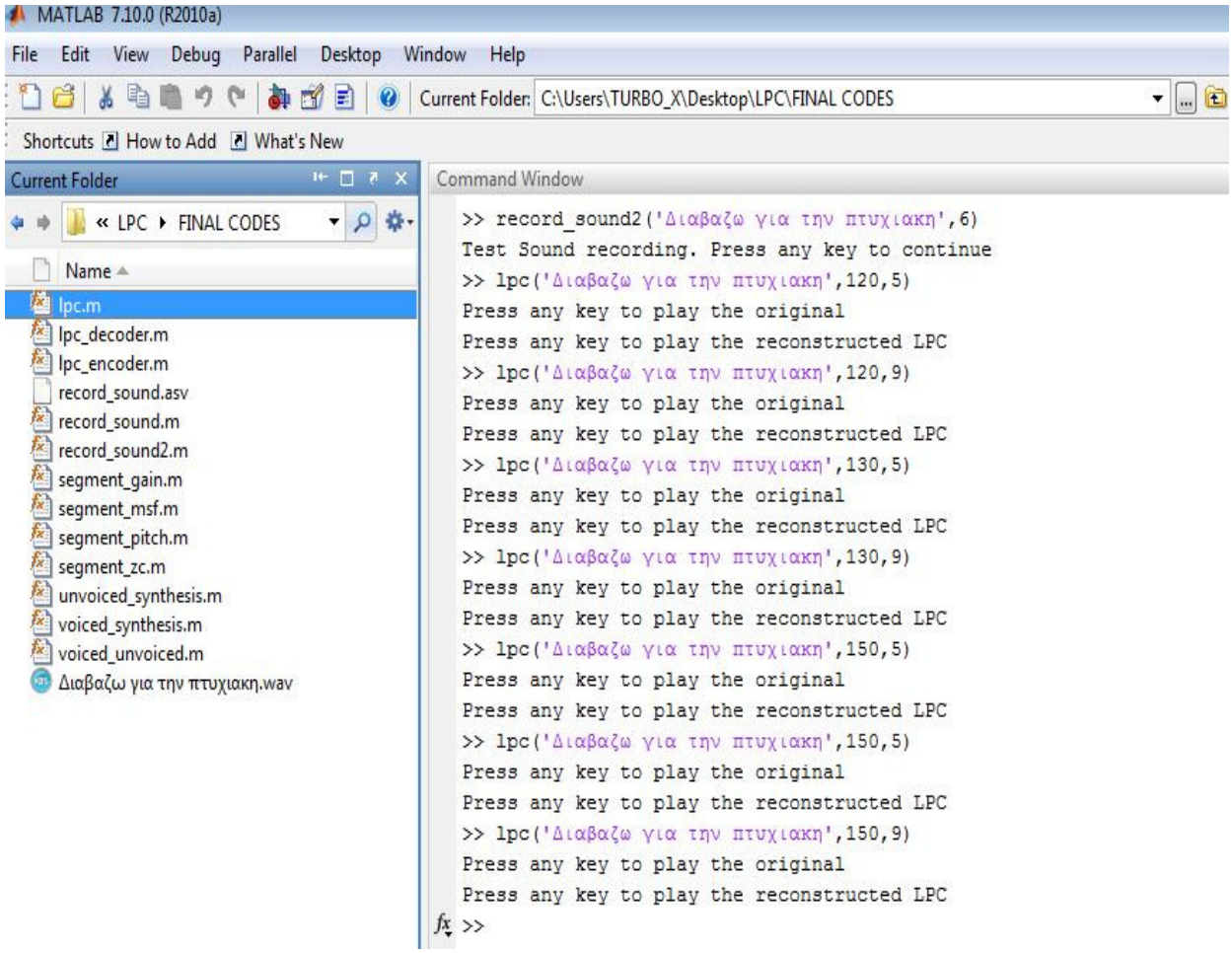
3.2 ΠΕΙΡΑΜΑΤΙΚΗ ΔΙΑΔΙΚΑΣΙΑ

Για την υλοποίηση του πειράματος ηχογραφήθηκαν 5 αρχεία ομιλίας. Στη συνέχεια το κάθε αρχείο ομιλίας εισήχθη στο πρόγραμμα και δοκιμάστηκε με 6 διαφορετικούς συνδυασμούς τμημάτων και συντελεστών. Στον παρακάτω πίνακα φαίνονται οι παράμετροι που χρησιμοποιήθηκαν.

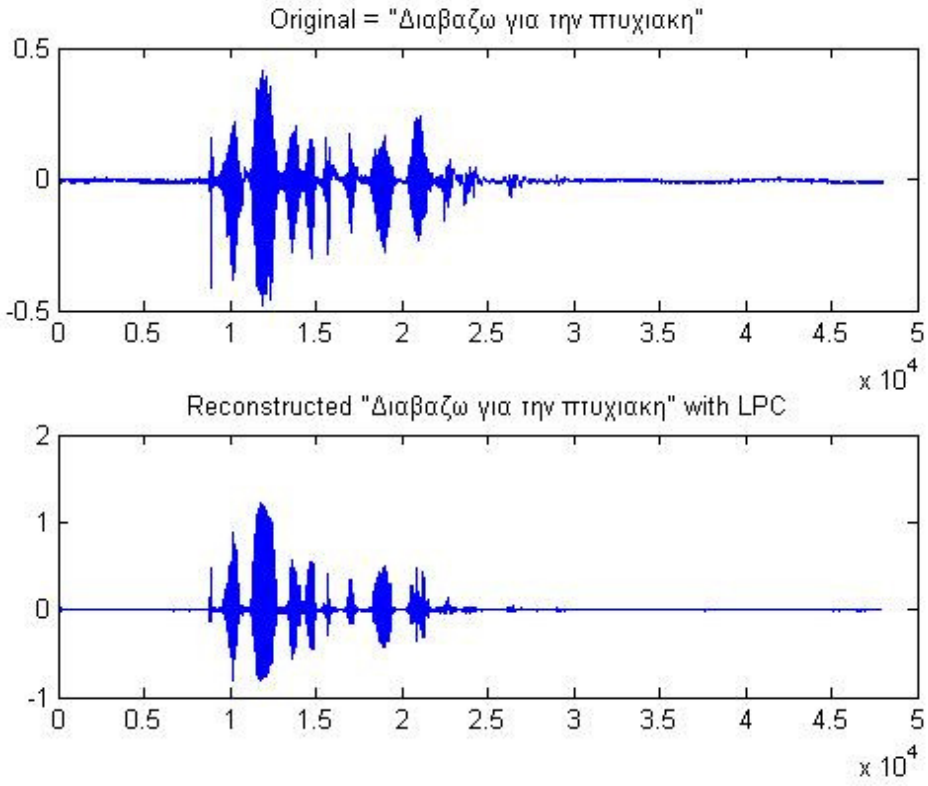
Πλήθος δειγμάτων τμήματος	Συντελεστές
120	5
120	9
130	5
130	9
150	5
150	9

Πίνακας 2 Παράμετροι πειράματος

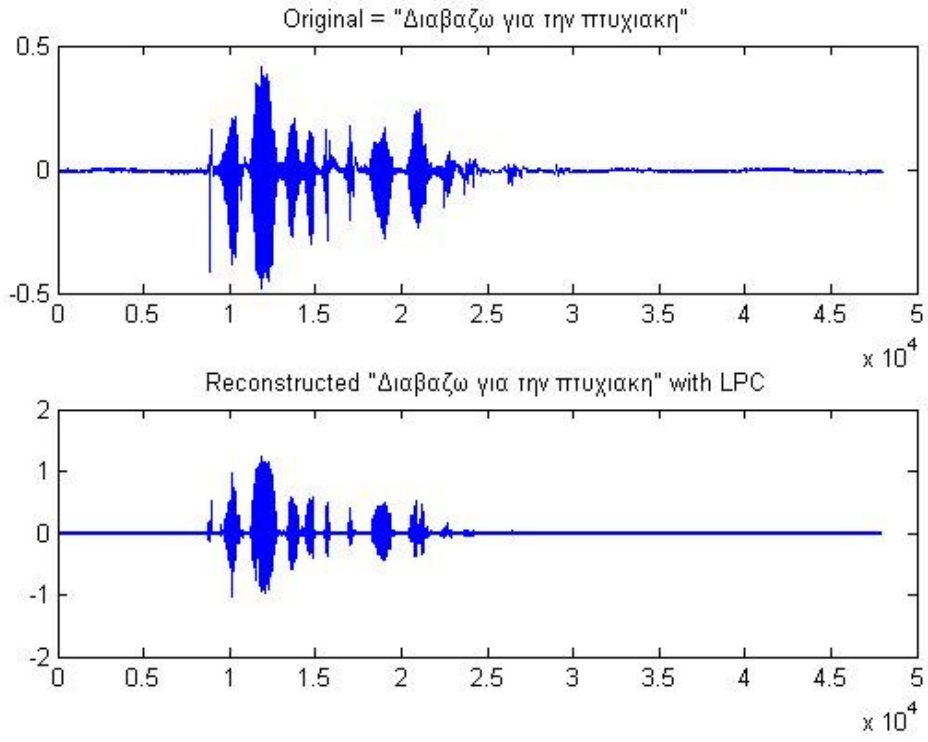
Η πειραματική διαδικασία ξεκινά με την ηχογράφηση της φράσης **"Διαβάζω για την πτυχιακή"** μέσω της συνάρτησης **record_sound2**. Στην συνέχεια καλείται η κύρια συνάρτηση **lpc** 6 φορές, μια για κάθε συνδυασμό παραμέτρων του παραπάνω πίνακα και εμφανίζονται οι αντίστοιχες γραφικές παραστάσεις.



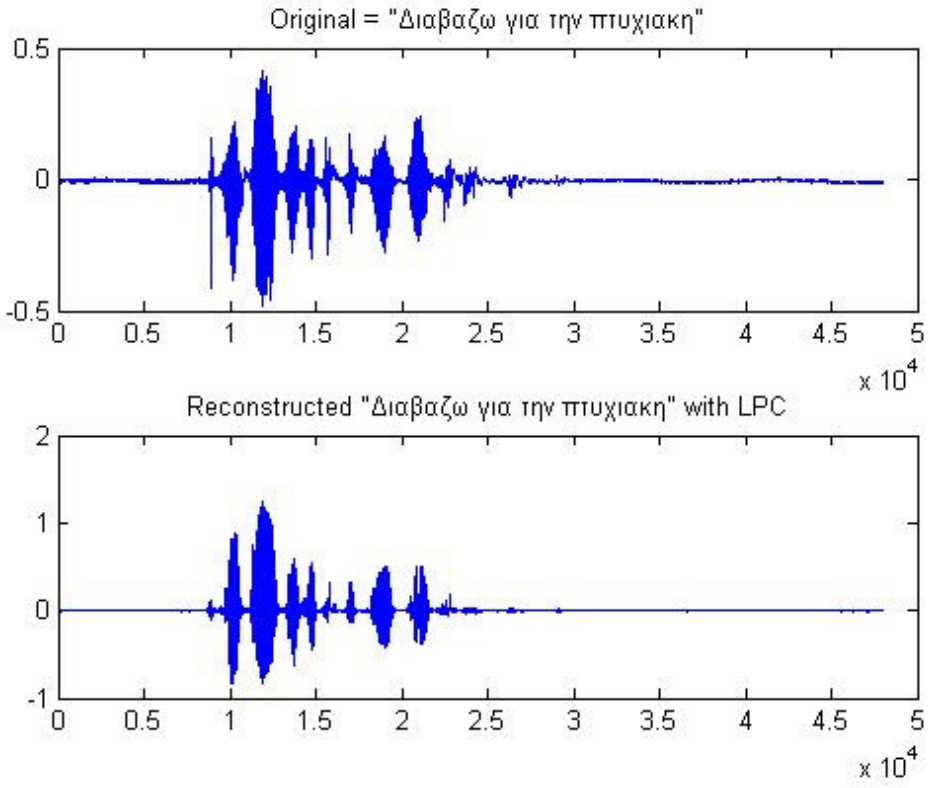
Εικόνα 3 Υλοποίηση της φράσης "Διαβάζω για την πτυχιακή" σε περιβάλλον matlab



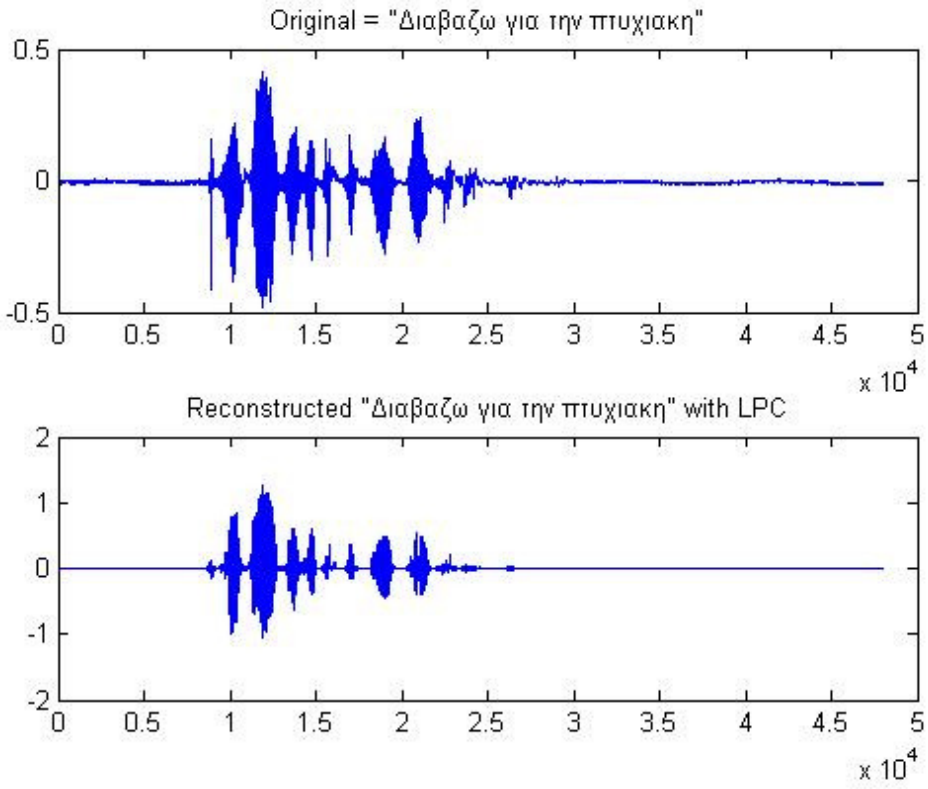
Εικόνα 4 Γραφική παράσταση του σήματος χωρισμένο σε 120 τμήματα και 5 συντελεστές



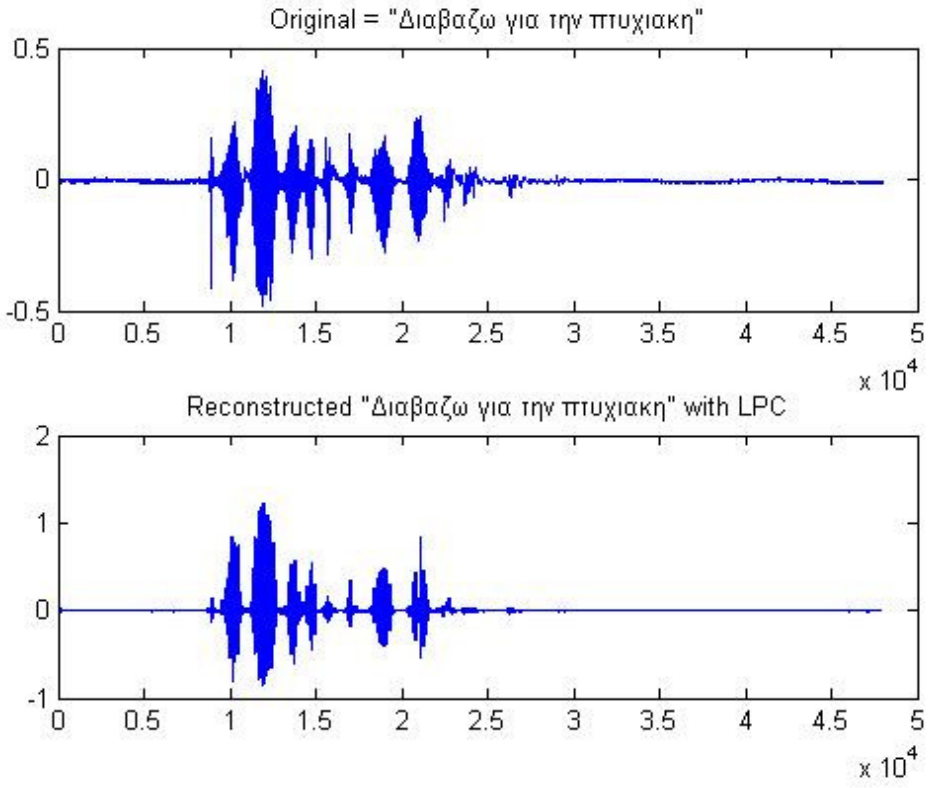
Εικόνα 5 Γραφική παράσταση του σήματος χωρισμένο σε 120 τμήματα και 9 συντελεστές



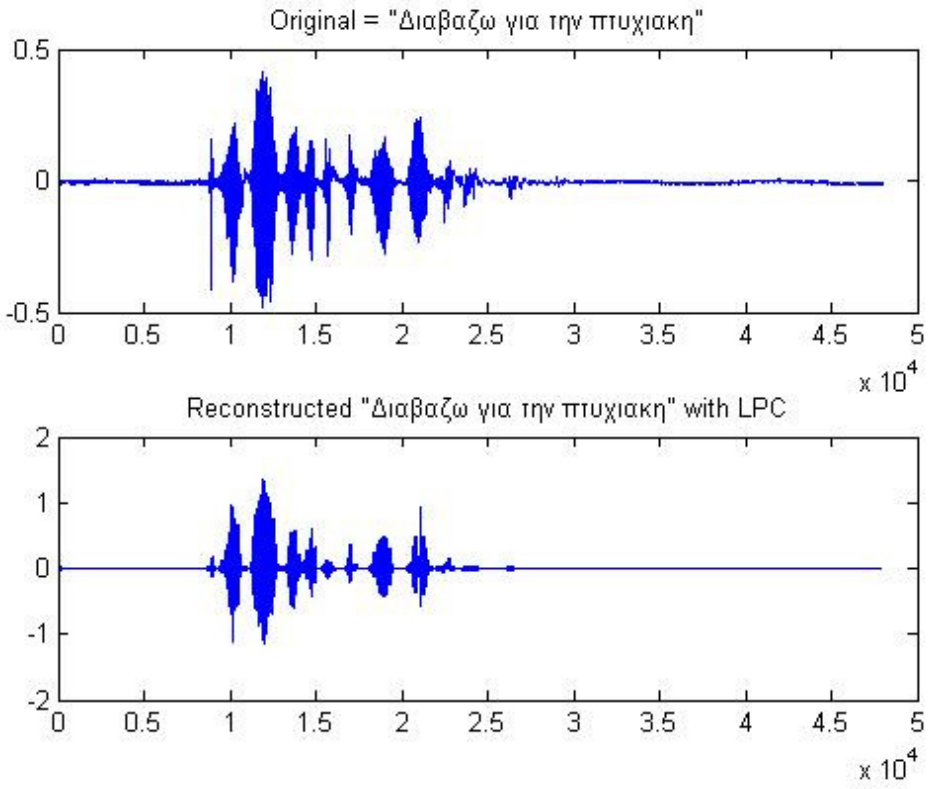
Εικόνα 6 Γραφική παράσταση του σήματος χωρισμένο σε 130 τμήματα και 5 συντελεστές



Εικόνα 7 Γραφική παράσταση του σήματος χωρισμένο σε 130 τμήματα και 9 συντελεστές

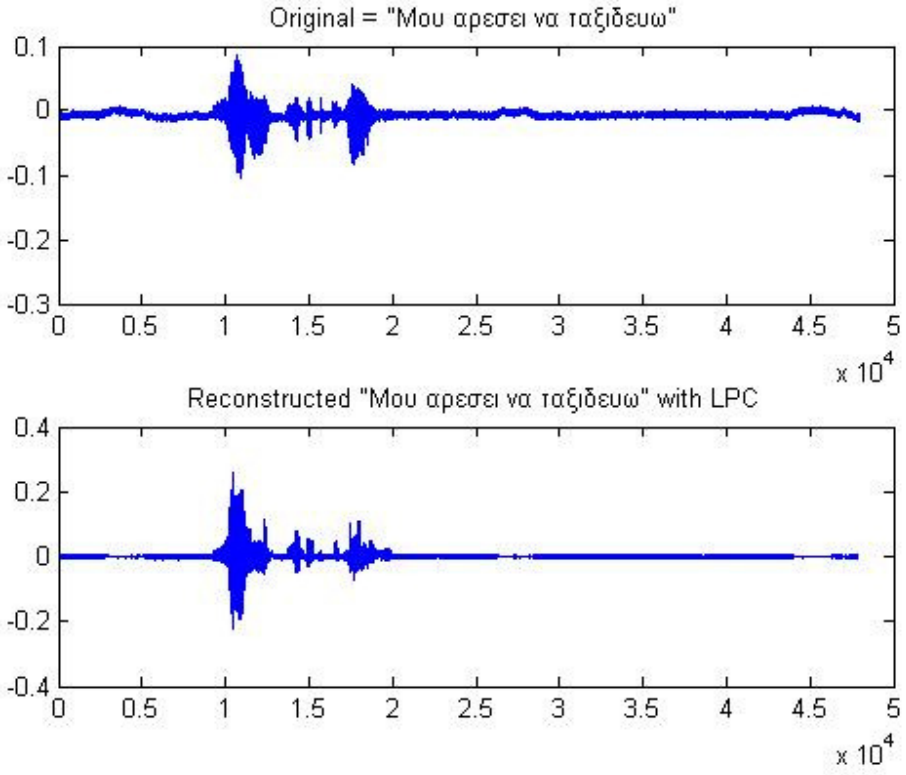


Εικόνα 8 Γραφική παράσταση του σήματος χωρισμένο σε 150 τμήματα και 5 συντελεστές

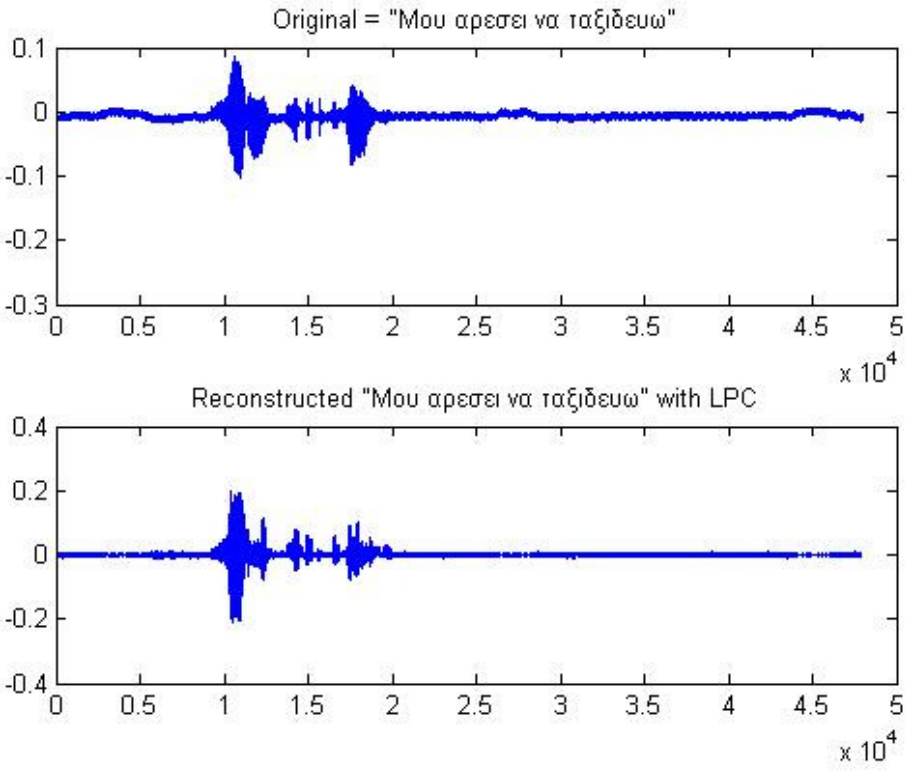


Εικόνα 9 Γραφική παράσταση του σήματος χωρισμένο σε 150 τμήματα και 9 συντελεστές

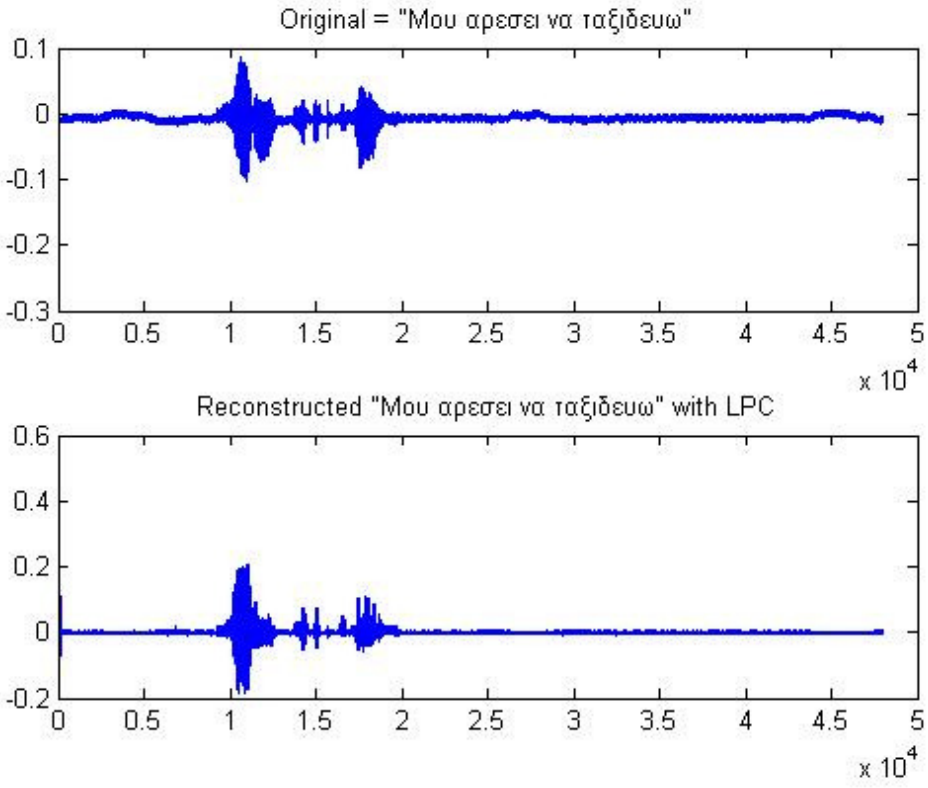
Στην συνέχεια ακολουθούν οι γραφικές παραστάσεις τεσσάρων ακόμα ηχογραφημένων μηνυμάτων αλλάζοντας κατά τον ίδιο τρόπο τις παραμέτρους.



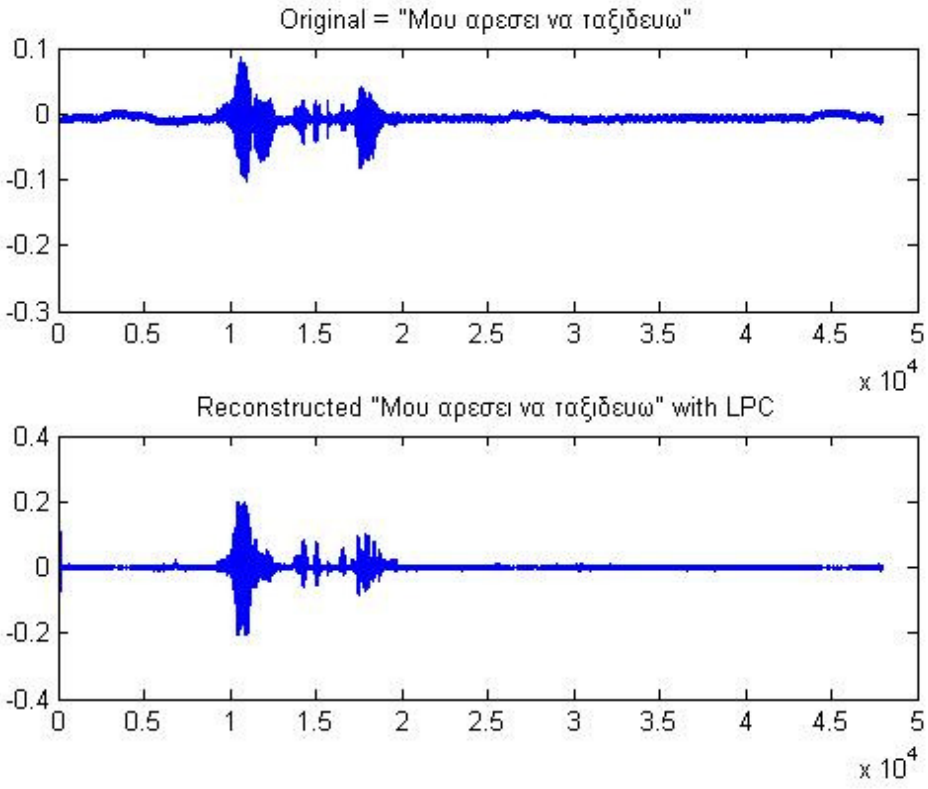
Εικόνα 10 Γραφική παράσταση του σήματος χωρισμένο σε 120 τμήματα και 5 συντελεστές



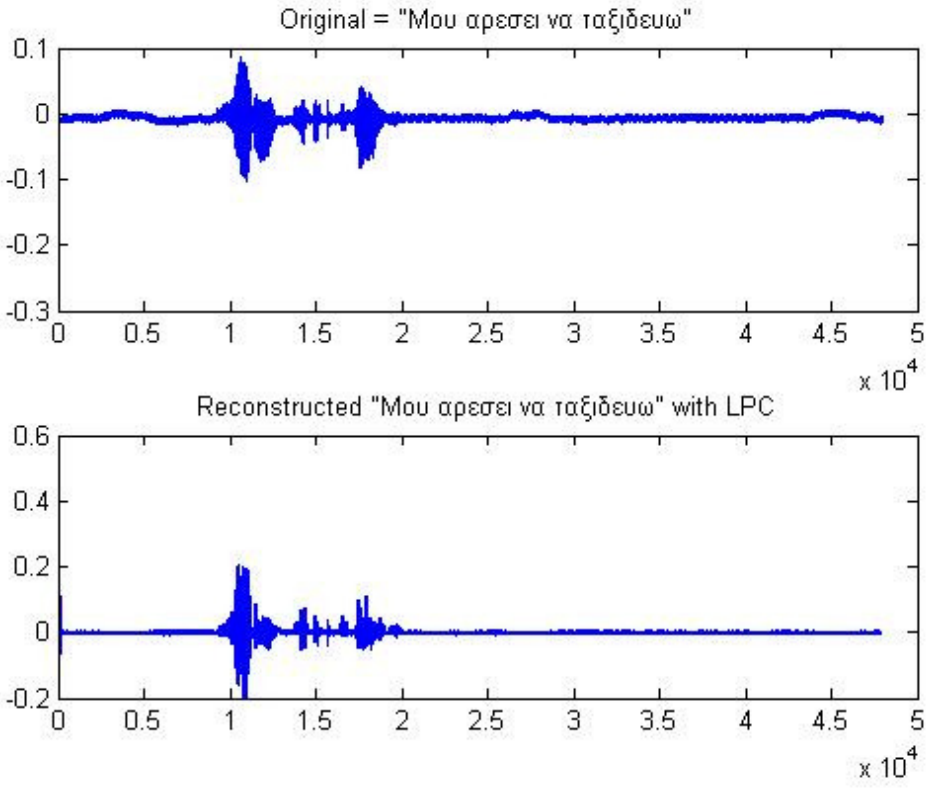
Εικόνα 11 Γραφική παράσταση του σήματος χωρισμένο σε 120 τμήματα και 9 συντελεστές



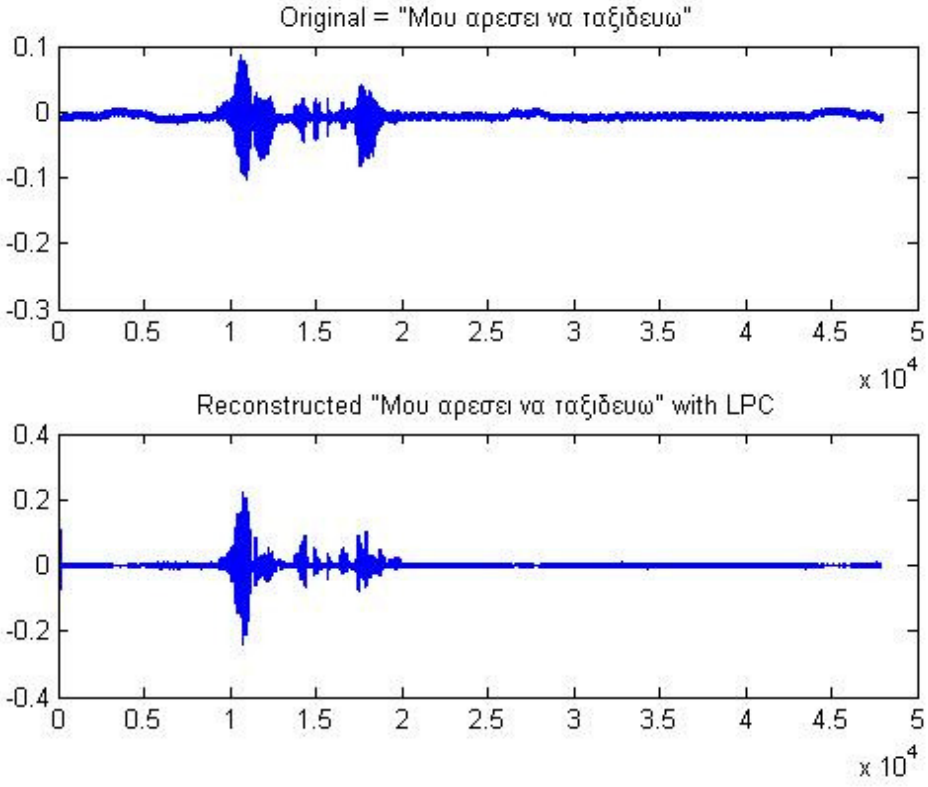
Εικόνα 12 Γραφική παράσταση του σήματος χωρισμένο σε 130 τμήματα και 5 συντελεστές



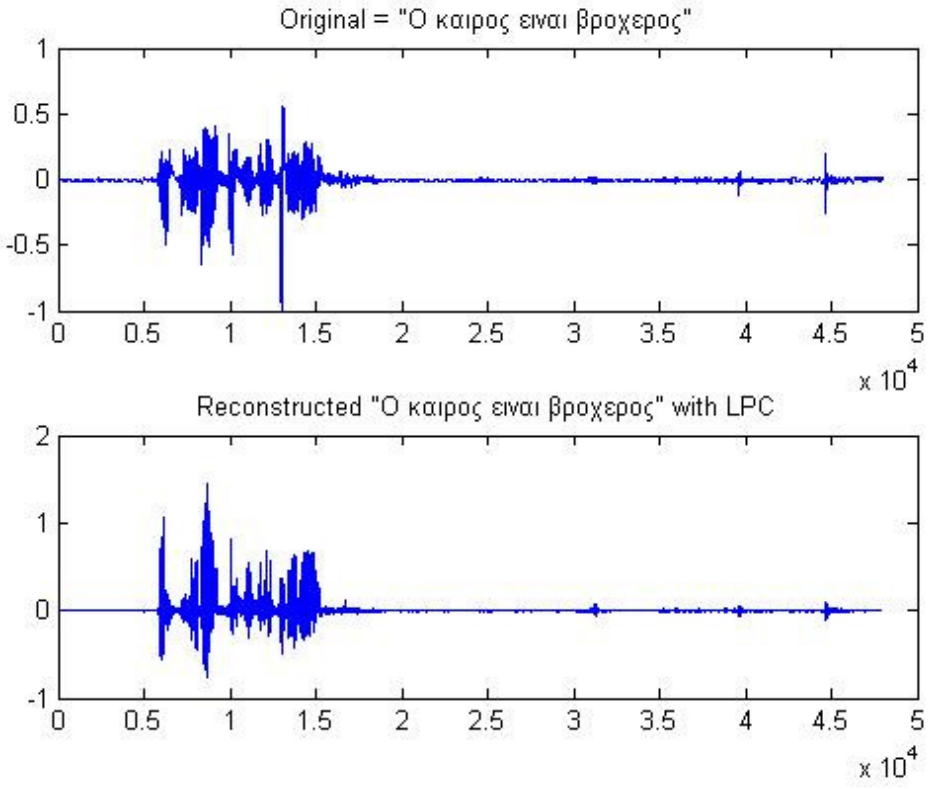
Εικόνα 13 Γραφική παράσταση του σήματος χωρισμένο σε 130 τμήματα και 9 συντελεστές



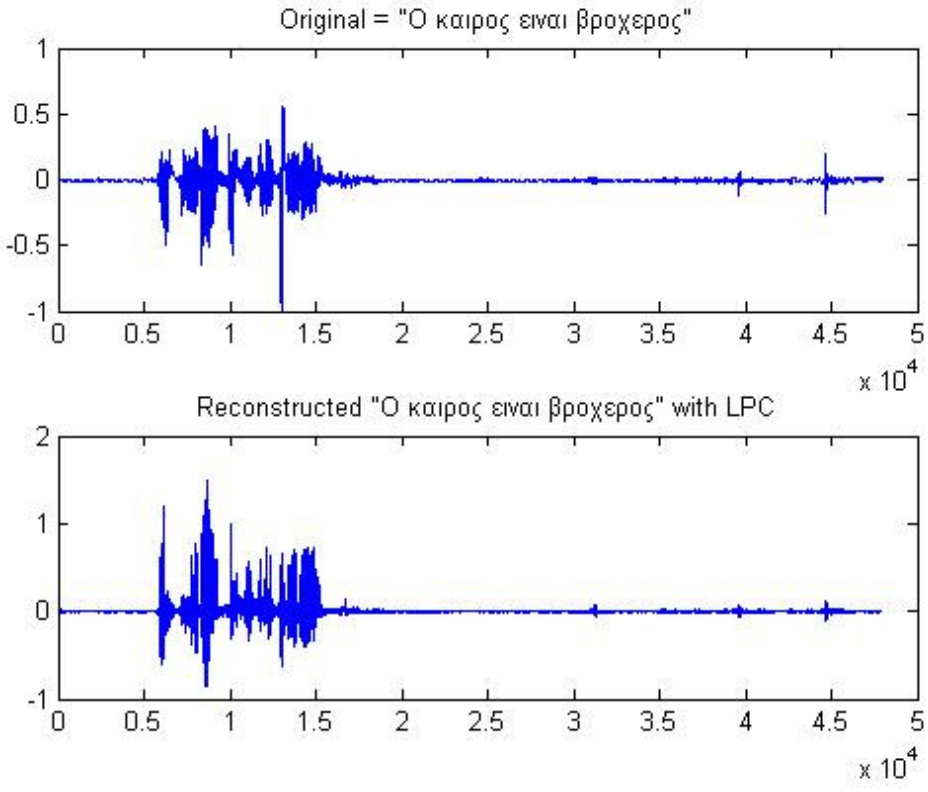
Εικόνα 14 Γραφική παράσταση του σήματος χωρισμένο σε 150 τμήματα και 5 συντελεστές



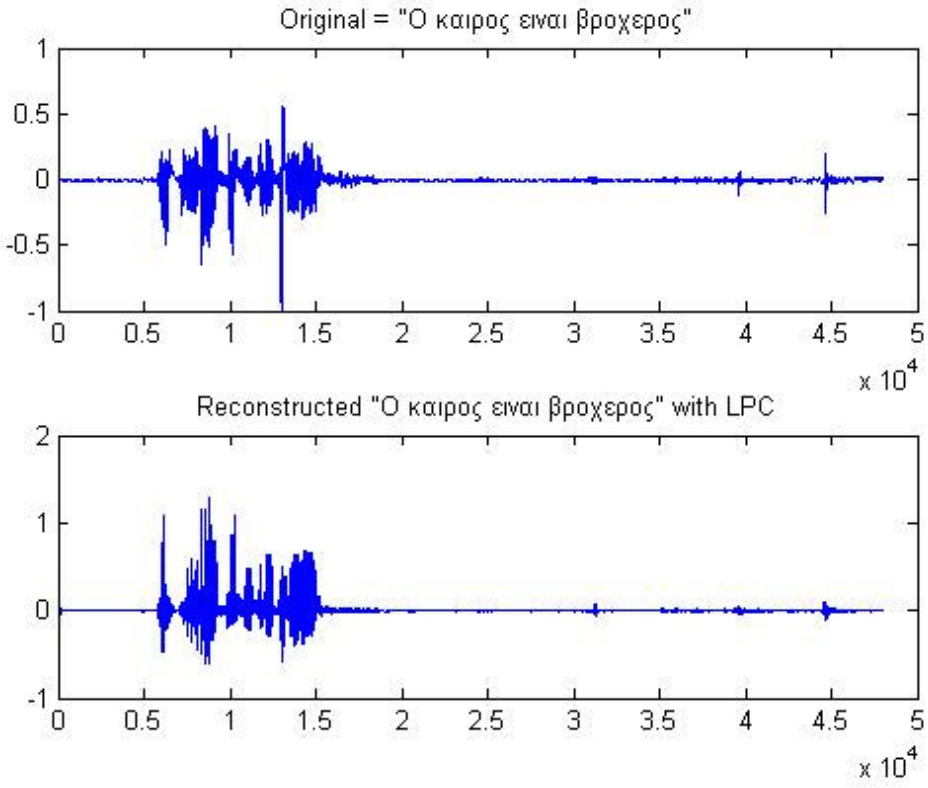
Εικόνα 15 Γραφική παράσταση του σήματος χωρισμένο σε 150 τμήματα και 9 συντελεστές



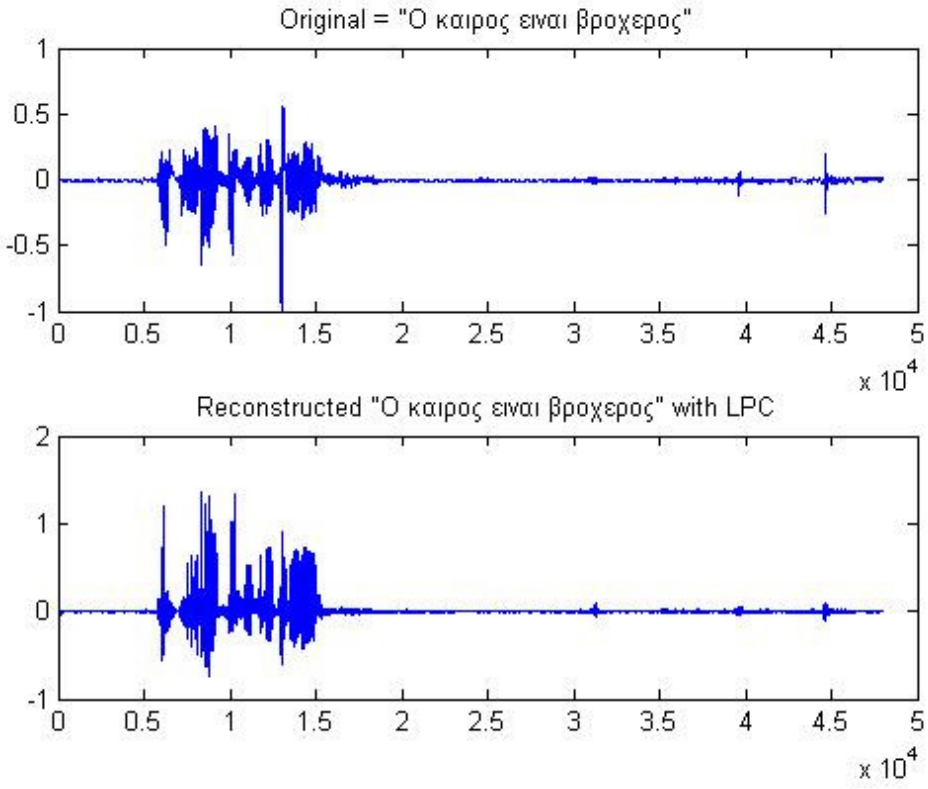
Εικόνα 16 Γραφική παράσταση του σήματος χωρισμένο σε 120 τμήματα και 5 συντελεστές



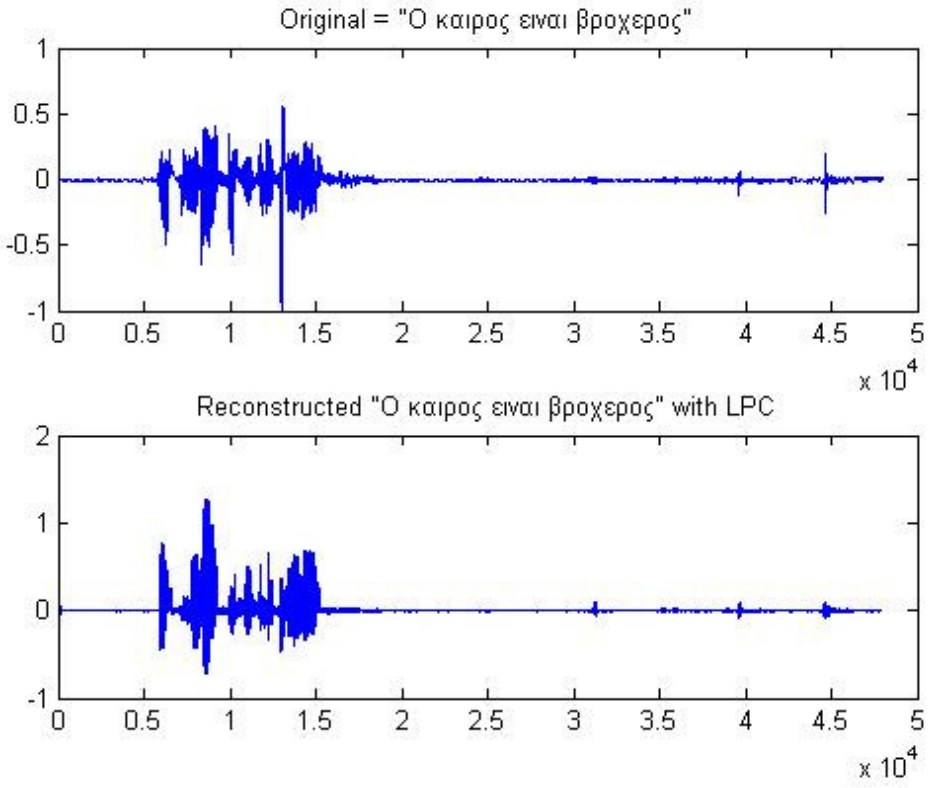
Εικόνα 17 Γραφική παράσταση του σήματος χωρισμένο σε 120 τμήματα και 9 συντελεστές



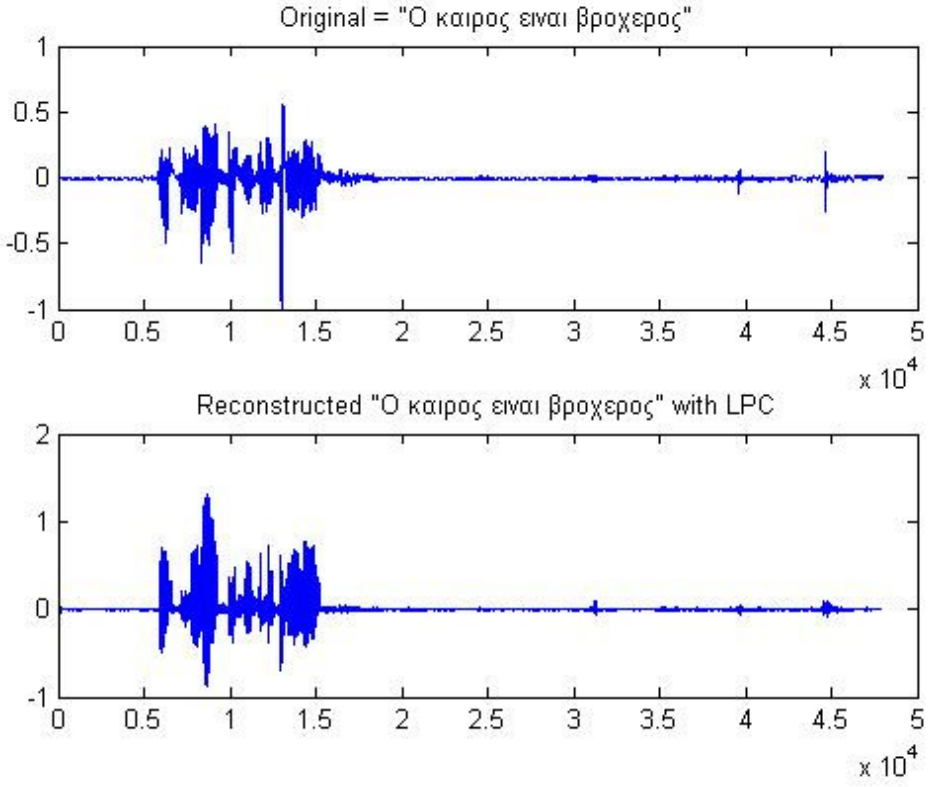
Εικόνα 18 Γραφική παράσταση του σήματος χωρισμένο σε 130 τμήματα και 5 συντελεστές



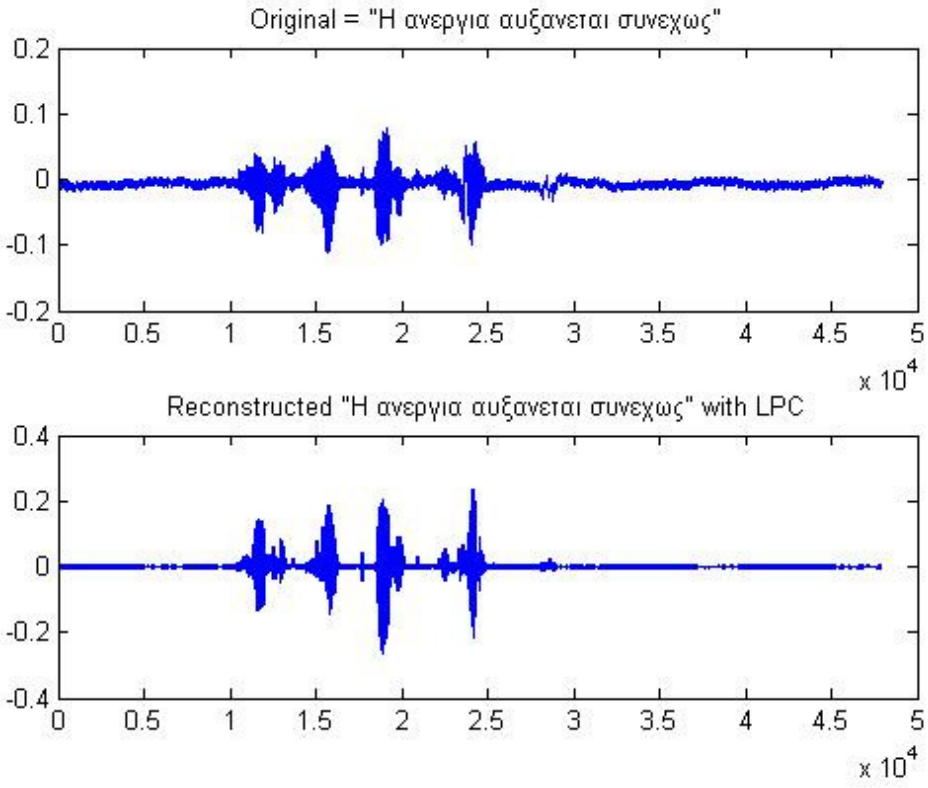
Εικόνα 19 Γραφική παράσταση του σήματος χωρισμένο σε 130 τμήματα και 9 συντελεστές



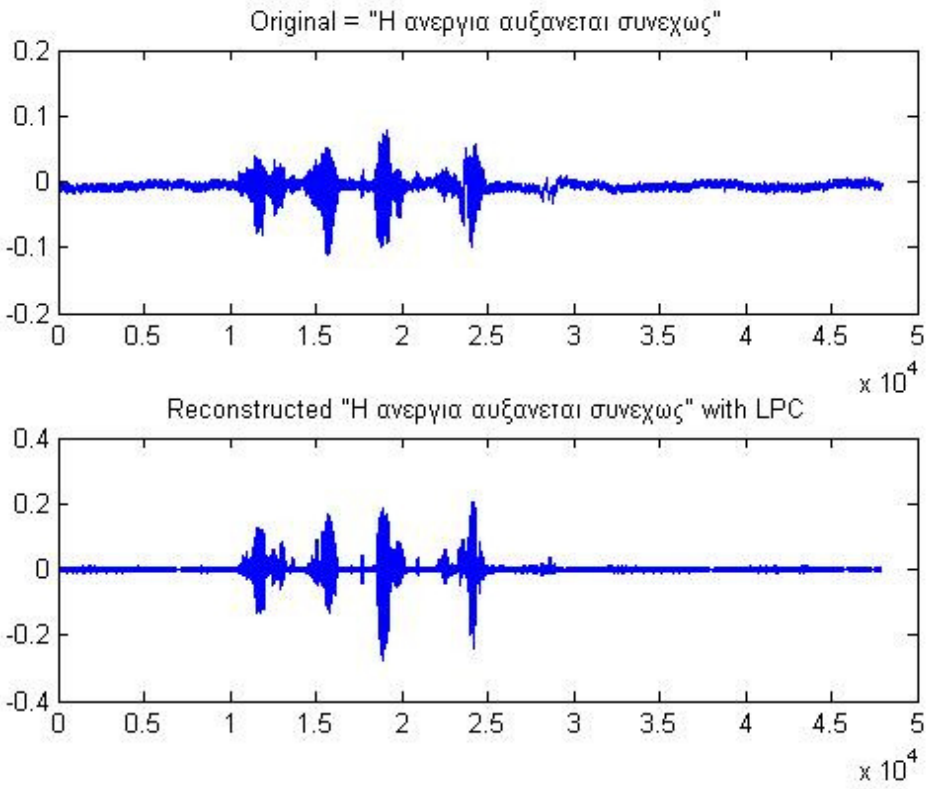
Εικόνα 20 Γραφική παράσταση του σήματος χωρισμένο σε 150 τμήματα και 5 συντελεστές



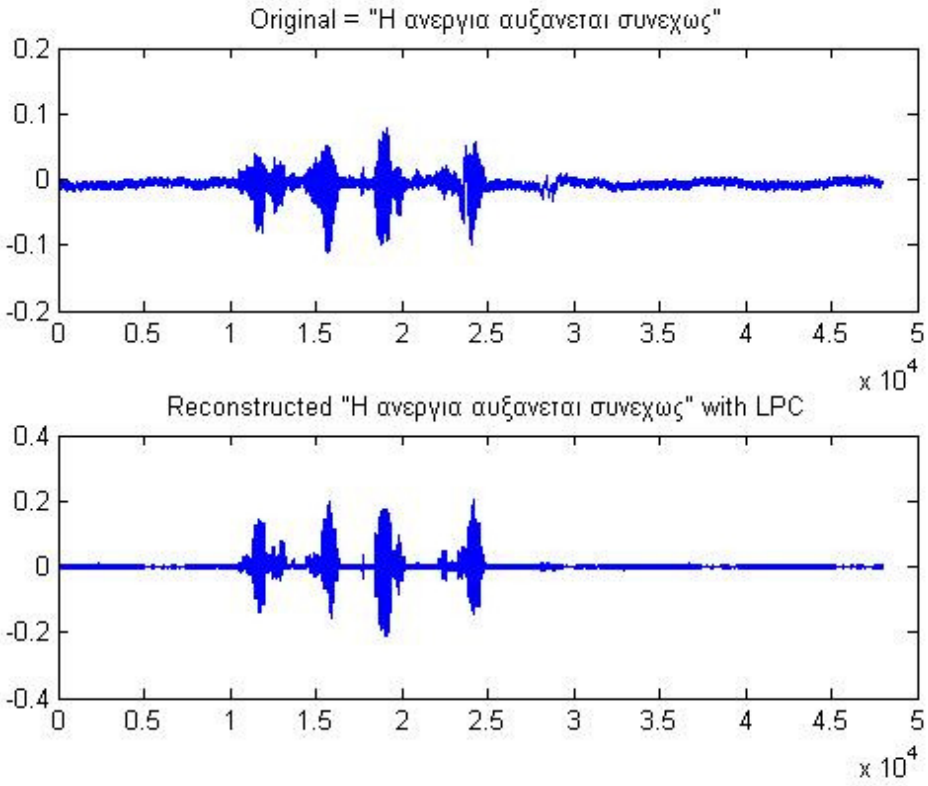
Εικόνα 21 Γραφική παράσταση του σήματος χωρισμένο σε 150 τμήματα και 9 συντελεστές



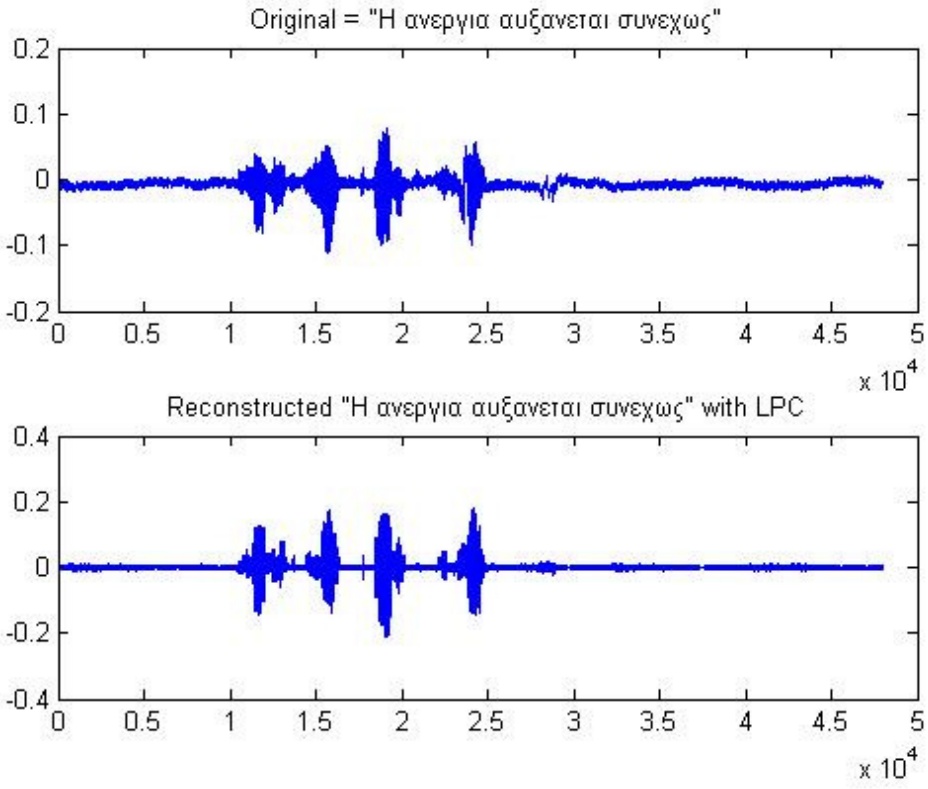
Εικόνα 22 Γραφική παράσταση του σήματος χωρισμένο σε 120 τμήματα και 5 συντελεστές



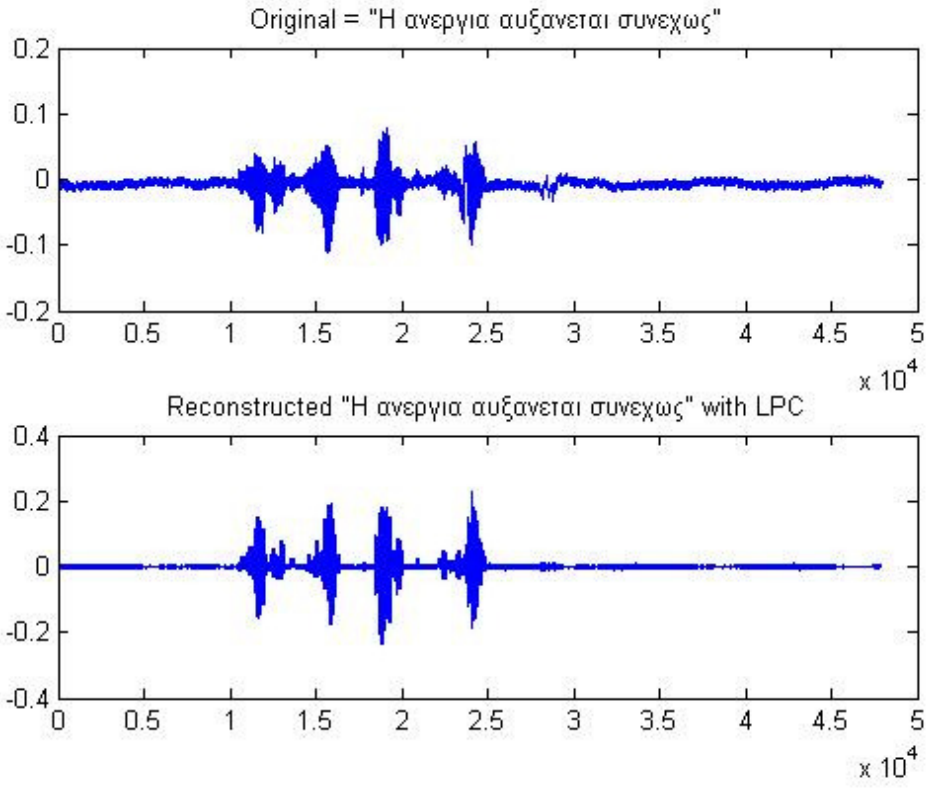
Εικόνα 23 Γραφική παράσταση του σήματος χωρισμένο σε 120 τμήματα και 9 συντελεστές



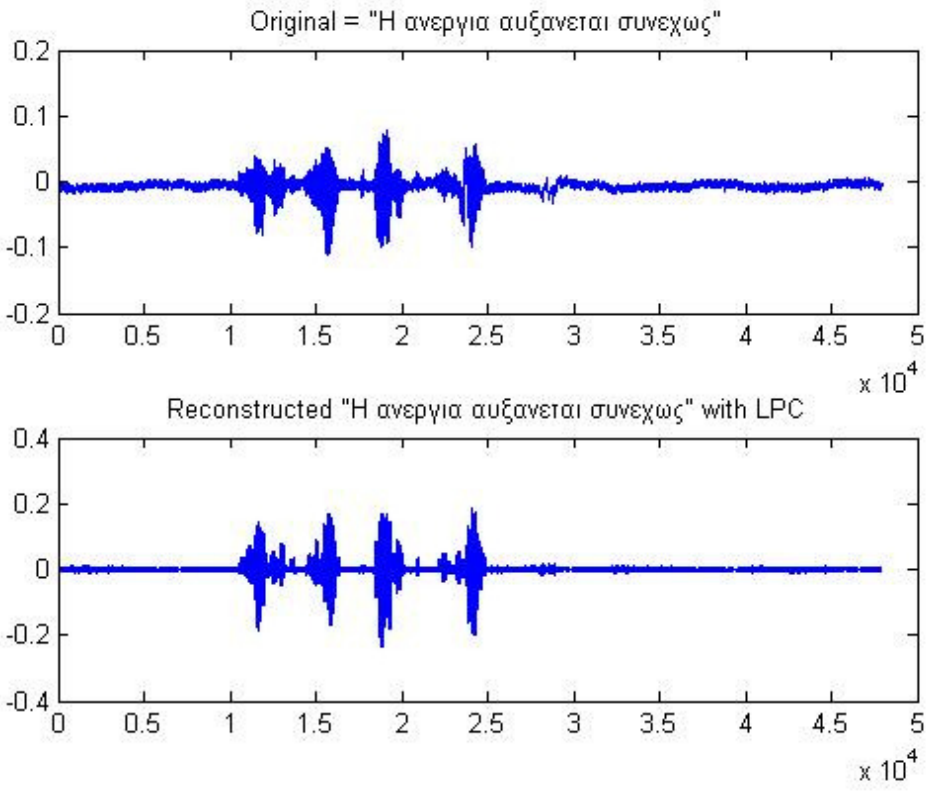
Εικόνα 24 Γραφική παράσταση του σήματος χωρισμένο σε 130 τμήματα και 5 συντελεστές



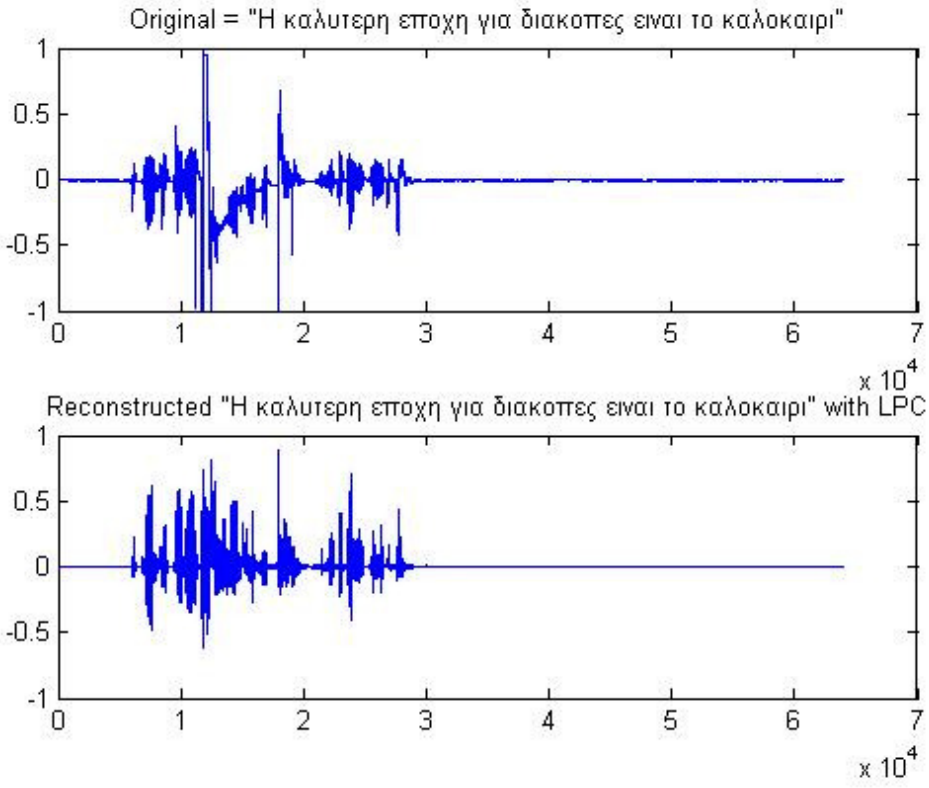
Εικόνα 25 Γραφική παράσταση του σήματος χωρισμένο σε 130 τμήματα και 9 συντελεστές



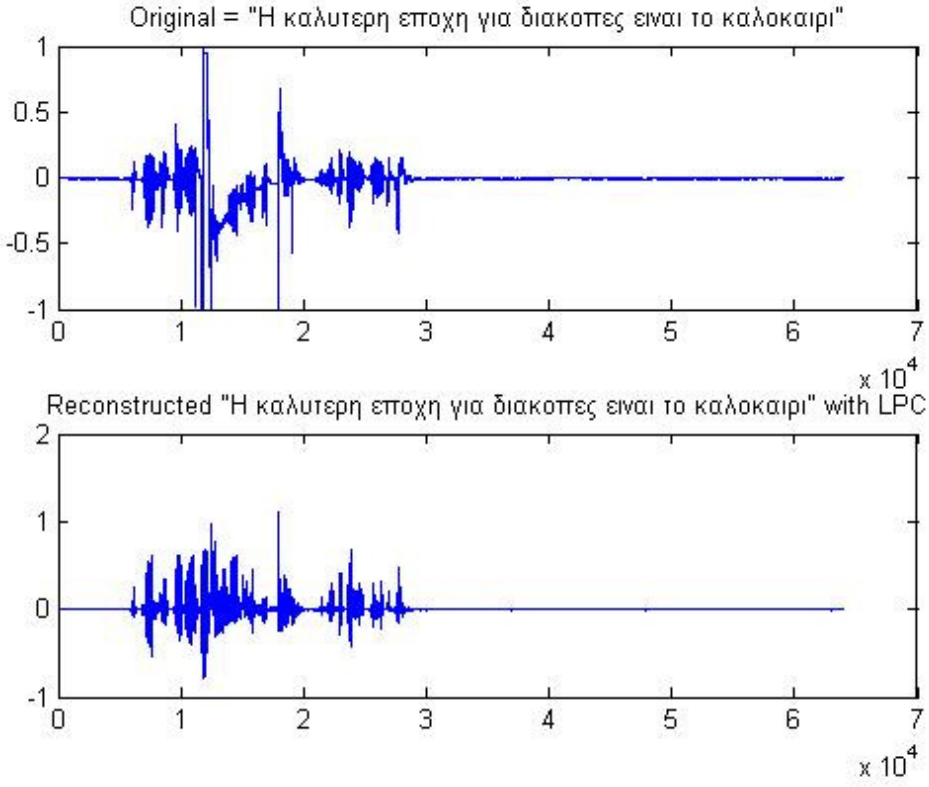
Εικόνα 26 Γραφική παράσταση του σήματος χωρισμένο σε 150 τμήματα και 5 συντελεστές



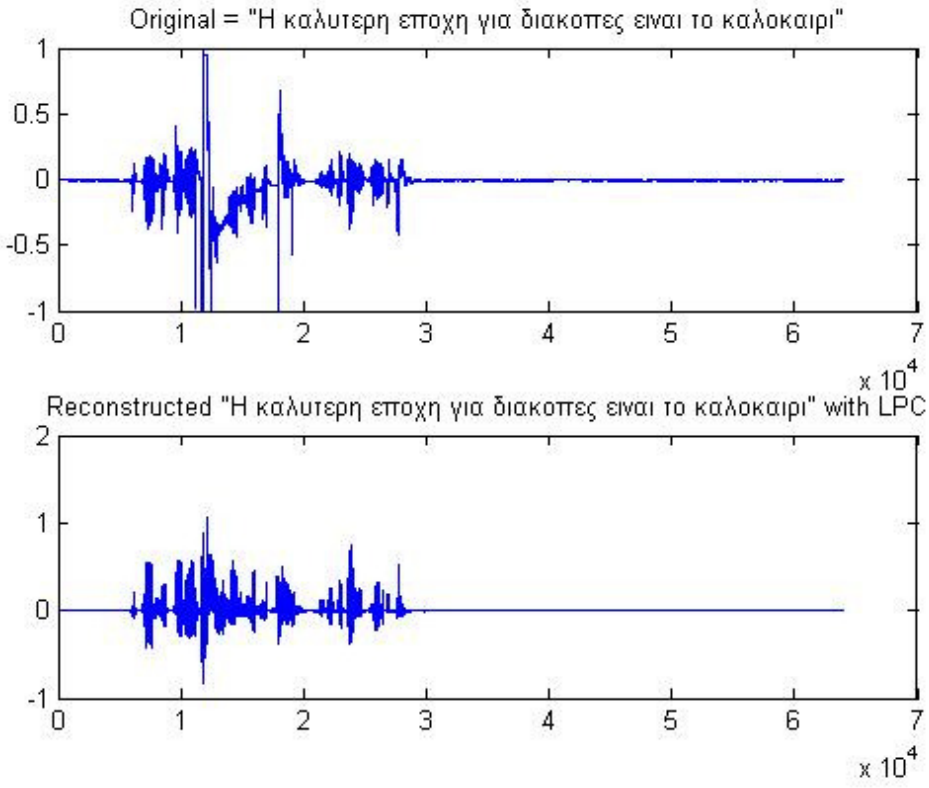
Εικόνα 27 Γραφική παράσταση του σήματος χωρισμένο σε 150 τμήματα και 9 συντελεστές



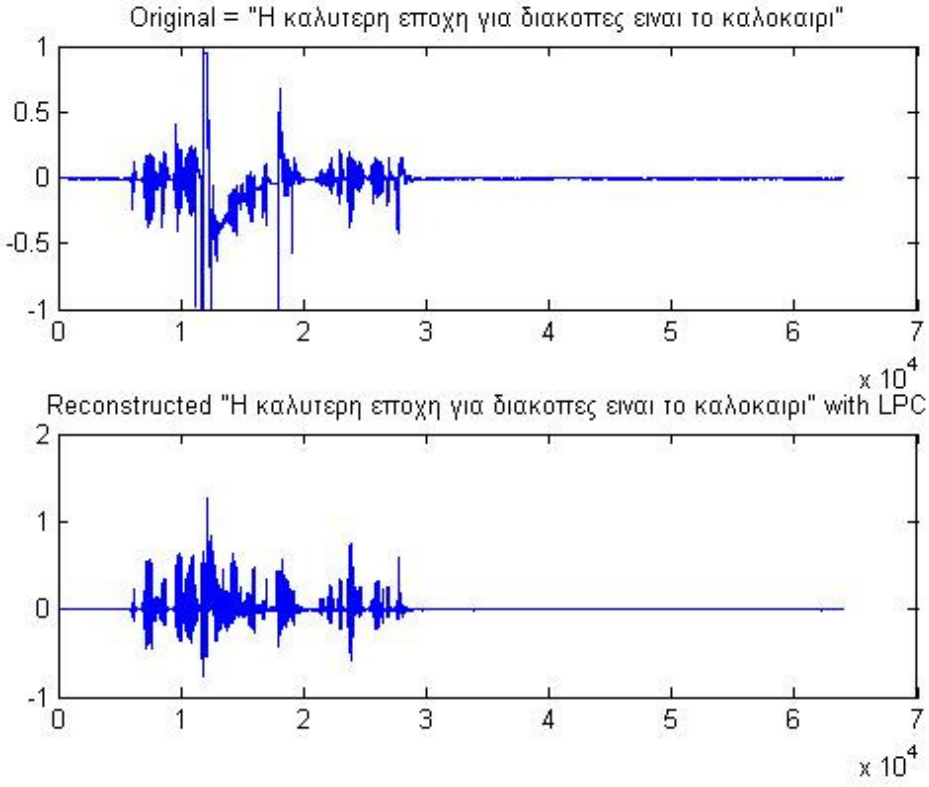
Εικόνα 28 Γραφική παράσταση του σήματος χωρισμένο σε 120 τμήματα και 5 συντελεστές



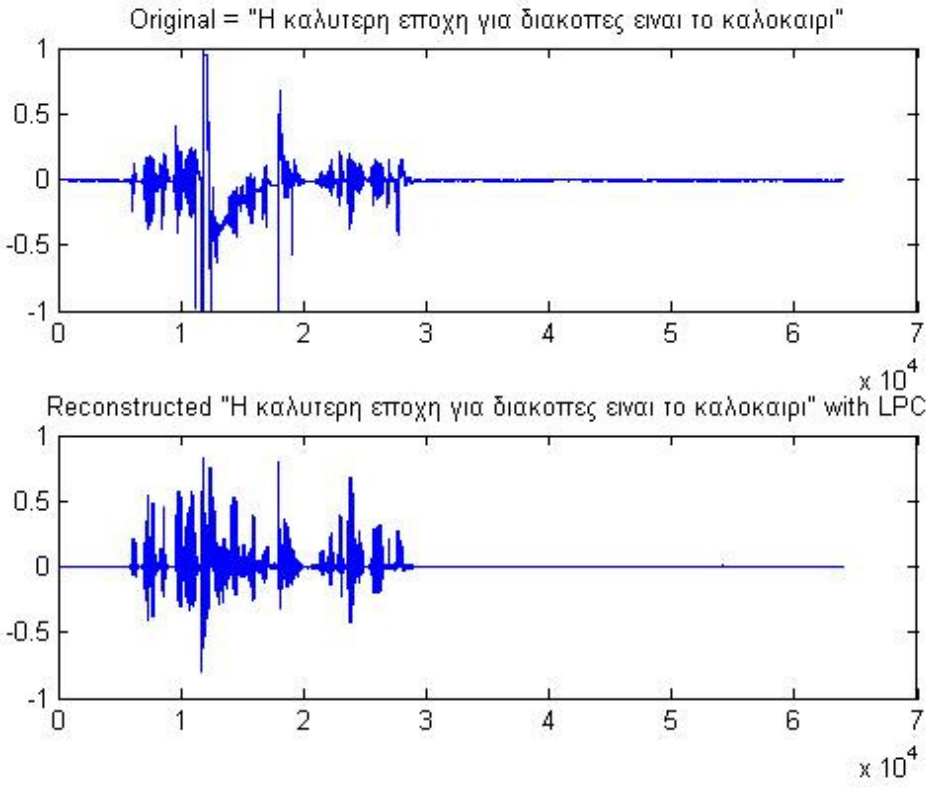
Εικόνα 29 Γραφική παράσταση του σήματος χωρισμένο σε 120 τμήματα και 9 συντελεστές



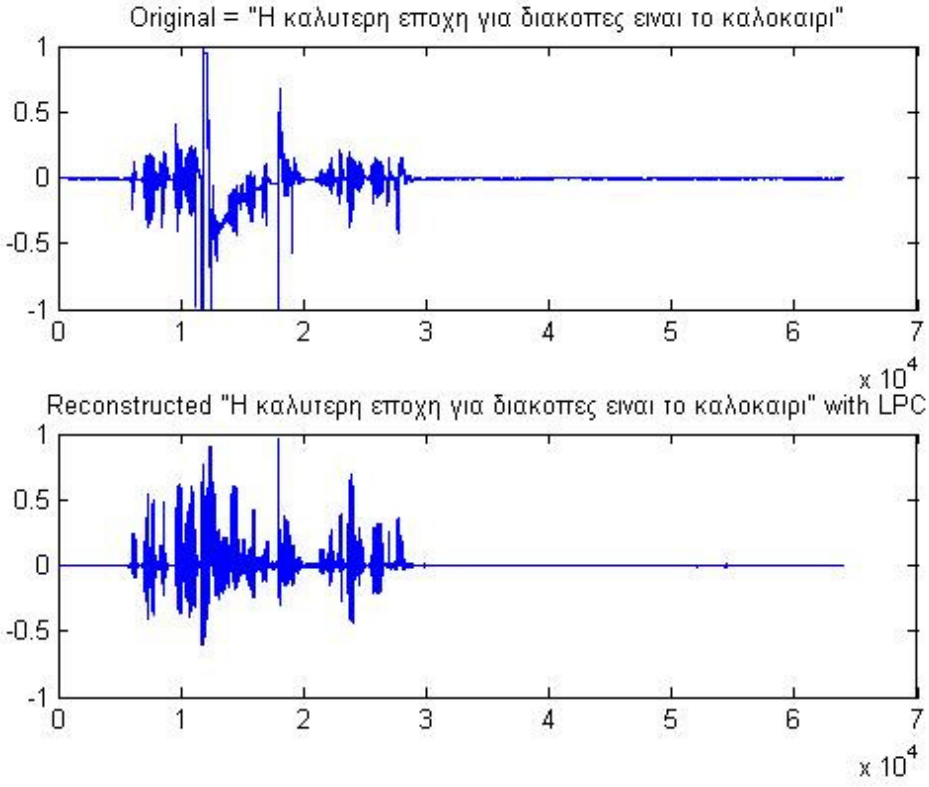
Εικόνα 30 Γραφική παράσταση του σήματος χωρισμένο σε 130 τμήματα και 5 συντελεστές



Εικόνα 31 Γραφική παράσταση του σήματος χωρισμένο σε 130 τμήματα και 9 συντελεστές



Εικόνα 32 Γραφική παράσταση του σήματος χωρισμένο σε 150 τμήματα και 5 συντελεστές



Εικόνα 33 Γραφική παράσταση του σήματος χωρισμένο σε 150 τμήματα και 9 συντελεστές

Συμπεράσματα

Η Γραμμική Κωδικοποίηση φωνής-Linear Predictive Coding (LPC) είναι μια τεχνική για ανάλυση και σύνθεση της ομιλίας με συμπίεση που επιχειρεί να μοντελοποιήσει την ανθρώπινη παραγωγή του ήχου, αντί να εκπέμπει μια εκτίμηση των κυμάτων του ήχου. Ο κωδικοποιητής του LPC διαχωρίζει ένα ηχητικό σήμα σε διαφορετικά τμήματα και στη συνέχεια στέλνει την πληροφορία για κάθε τμήμα προς τον αποκωδικοποιητή. Ο κωδικοποιητής στέλνει πληροφορίες σχετικά με το εάν ένα τμήμα έχει ήχο ή όχι και την περίοδο **pitch** για το ηχηρό τμήμα που χρησιμοποιείται για να δημιουργήσει το αντίστοιχο σήμα στο αποκωδικοποιητή. Ο κωδικοποιητής στέλνει επίσης πληροφορία σχετικά με τη φωνητική οδό η οποία χρησιμοποιείται για την κατασκευή ενός φίλτρου στην πλευρά του αποκωδικοποιητή το οποίο όταν δοθεί το σήμα ως είσοδο μπορεί να αναπαράγει την αρχική ομιλία. Το κύριο πλεονέκτημα είναι ότι η διαδικασία κωδικοποίησης και αποκωδικοποίησης της φωνής δίνεται από ένα απλοποιημένο μοντέλο και την αναλογία ενός μοντέλου πηγής-φίλτρου με το σύστημα παραγωγής της ομιλίας. Είναι μια χρήσιμη μέθοδος για την κωδικοποίηση ομιλίας σε χαμηλό ρυθμό bit. Ωστόσο η απόδοση του LPC περιορίζεται από την ίδια τη μέθοδο και τα τοπικά χαρακτηριστικά του σήματος. Ένα σωστό μοντέλο all-pole για το φάσμα του σήματος είναι δύσκολο να βρεθεί. Δεν αντιπροσωπεύει την επιθυμητή φασματική πληροφορία που πρέπει να διαμορφωθεί, αφού ενδιαφέρομαστε για την τοποθέτηση της φασματικής περιβάλλουσας όσο το δυνατόν πλησιέστερα και όχι οσον αφορα το αρχικό φάσμα. Η φασματική περιβάλλουσα πρέπει να είναι μια ομαλή λειτουργία που να διέρχεται μέσα στις εξέχουσες κορυφές του φάσματος , αποδίδοντας μια επίπεδη αλληλουχία, και όχι τις «κοιλάδες» που σχηματίζονται από τις κορυφές αρμονικών .

Εκτελώντας τη μέθοδο LPC διαπιστώνουμε ότι το ανακατασκευασμένο σήμα που δημιουργείται είναι παρόμοιο με το αρχικό όσων αφορά το πλάτος και τις διακυμάνσεις. Για τα διάφορα δείγματα φωνής που εισάγαμε στην υλοποιημένη μέθοδο LPC διαπιστώνουμε ότι ποιοτικά το ανακατασκευασμένο σήμα φωνής είναι μια αλλοιωμένη μορφή του αρχικού. Από τα πειράματα μας διαπιστώσαμε ότι η αύξηση του αριθμού των τμημάτων και συντελεστών δεν αναβαθμίζει ποιοτικά όπως θα περιμέναμε θεωρητικά. Με την άυξηση των τμημάτων θεωρητικά περιμένουμε βελτίωση του ανακατασκευασμένου τμήματος αφού είναι περισσότερο ευδιάκριτες οι περιοχές των ηχηρών και των άηχων περιοχών.

Από την άλλη μεριά η αύξηση των συντελεστών με αποτελεί κρητίριο για την ποιοτική βελτίωση αφού μπορεί να έχουν περιτπώσεις overfitting στην διαδικασία ανκατασκευής του σήματος.

Σε αυτήν την διπλωματική πραγματευθήκαμε μόνο με την ποιοτική σύγκριση και όχι με την ποσοτική δηλαδή σύγκριση των σημάτων του αρχικού και του ανακατασκευασμένου σήματος με μετρικές όπως αυτή του μέσου τετραγωνικού σχφάλματος (MSE) το οποίο όμως μπορέι να είναι και παραπλανητικό καθώς μπορεί από ένα peak του ανακατασκευασμένου σηματος να δώσει πολύ μεγάλη τιμή.

Αναφορές

[1] http://nefeli.lib.teicrete.gr/browse2/stef/thl/2006/Aggelis/attached-document/2006Aggelis.pdf
[2] http://www.teiser.gr/icd_old/ptixiakes_parousiaseis/pipis.pdf
[3] Rabiner, L. and Schafer, R.. Digital Processing of Speech Signals. Prentice-Hall.
[4] Kondoz, A.. Digital Speech - Coding for Low Bit Rate Communications Systems. Wiley. 1994.
[5] IEEE Communications Magazine. Special issue on Standardization and Characterization of G.729. Sep. 1997.
[6] Campbell, J. and Tremain, T.. Voiced / Unvoiced Classification of Speech with Applications to theU.S. Governement LPC-10E Algorithm. IEEE ICASSP. 1996. pp. 473-476.
[7] Ghaemmaghami, S.; Deriche, M.; Boashash, B. Edited by: Deriche, M.; Moody, M.; Bennamoun, M. A new approach to pitch and voicing detection through spectrum periodicity measurement. vol.2 , TENCON '97 Brisbane - Australia. Proceedings of IEEE TENCON '97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications Dec. 1997. New York, NY, USA.
[8] Juhar, J. Advanced pitch detection algorithms. DSP '97. 3rd International Conference on Digital Signal Processing. Proceedings of the Conference, Proceedings of Digital Signal Processing '97, Herl'any, Slovakia, 3-4 Sept. 1997. Kosice, Slovakia.
[9] Janer, L.; Bonet, J.J.; Lleida-Solano, E. Pitch detection and voiced/unvoiced decision algorithm based on wavelet transforms. vol.2 , Proceedings ICSLP 96. Fourth International Conference on Spoken Language Processing Cat. No.96TH8206 , Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96, Philadelphia, PA, USA, 3-6 Oct. 1996. New York, NY, USA.
[10]. J.D. Markel and A.H. Gray (Jr.), “Linear Prediction of Speech”, Springer-Verlag Berlin Heidelberg New York, New York, 1976, ISBN: 0-387-07563-1, pp 1-15.
[11]. John Holmes and Wendy Holmes, “Speech Synthesis and Recognition”, Second Edition, Taylor & Francis Group, New York, 2001, ISBN: 0-7484-0856-8, Ch. 1,4 and 6.
[12]. Richard L. Klevans and Robert D. Rodman, “Voice Recognition”, Artech House, Inc., MA, 1997, ISBN: 0-890006-927-1, pp 15-31.
[13]. Jerry D. Gibson, Toby Berger, Tom Lookabaugh, Dave Lindbergh and Richard L. Baker, “Digital Compression for Multimedia- Principles and Standards”, Morgan Kaufmann Publishers Inc., CA, 1998, ISBN: 1-55860-369-7, pp 1-4 & Ch. 6.
[14]. Nagarajan, S. Sankar, “Efficient implementation of linear predictive coding algorithms”, Proceedings of the IEEE, Southeastcon 1998, pp 69 – 72.
[15]. Matthew Hutchinson, “How speech can be modeled as a source signal passing through a filter.”, Dec. 2004 http://cnx.org/content/m12470/latest/ (Date accessed: 04/21/2010, 05/01/2010)
[14]. Uzdy, Z. , “Human speaker recognition performance of LPC voice processors”, IEEE Transactions on Acoustics, Speech and Signal processing, Vol. 38, Issue 12, 1998
[15]. Yedlapalli, S.S., “Transforming Real Linear Prediction Coefficients to Line Spectral Representations With a Real FFT”, IEEE Transactions on speech and audio processing, Vol. 13, Issue 6, 200570