Α.Τ.Ε.Ι. ΚΑΛΑΜΑΤΑΣ
ΠΑΡΑΡΤΗΜΑ ΣΠΑΡΤΗΣ
ΤΜΗΜΑ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΚΩΔΙΚΟΠΟΙΗΣΗ ΦΩΝΗΣ – ΚΩΔΙΚΟΠΟΙΗΤΕΣ ΧΑΜΗΛΟΥ ΡΥΘΜΟΥ ΜΕΤΑΔΟΣΗΣ

ΣΠΟΥΔΑΣΤΡΙΑ:
Ρενάτε Δημητρουλάκου
ΥΠΕΥΘΥΝΟΣ ΚΑΘΗΓΗΤΗΣ:
Μποζαντζής Βασίλειος

ΣΠΑΡΤΗ/ΜΑΙΟΣ 2012

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΙΝΑΚΑΣ ΕΙΚΟΝΩΝ

ΕΙΣΑΓΩΓΗ

Η πτυχιακή αυτή έχει θέμα «ΚΩΔΙΚΟΠΟΙΗΣΗ ΦΩΝΗΣ-ΚΩΔΙΚΟΠΟΙΗΤΕΣ ΧΑΜΗΛΟΥ ΡΥΘΜΟΥ ΜΕΤΑΔΟΣΗΣ». Στην εργασία μου θα σας παρουσιάσω μια μελέτη γύρω από την κωδικοποίηση της φωνής με τη χρήση των κωδικοποιητών φωνής χαμηλού ρυθμού μετάδοσης, ειδικότερα με τη χρήση των vocoder.

Η εργασία χωρίζεται σε τρία βασικά μέρη. Το πρώτο μέρος περιλαμβάνει το 1 κεφάλαιο όπου εκεί θα δούμε τον τρόπο παραγωγής της φωνής, ορισμένα βασικά χαρακτηριστικά της φωνής και θα πάρουμε μια συνοπτική ιδέα των κωδικοποιητών που έχουμε στη διάθεση μας αλλά και τους τρόπους με τους οποίους τους αξιολογούμε.

Στο δεύτερος μέρος συμπεριλαμβάνονται τα κεφάλαια 2, 3, 4 και 5. Σε αυτό το μέρος θα δούμε αναλυτικά τις τεχνικές και τους κωδικοποιητές που χρησιμοποιούνται για την κωδικοποίηση φωνής χαμηλού ρυθμού μετάδοσης. Η μελέτη μου αρχίζει με το LPC μοντέλο και καταλήγει με δύο κωδικοποιητές με ιδιαίτερα χαμηλό ρυθμό μετάδοσης.

Στο τελευταίο κεφάλαιο 6 παρουσιάζονται η υλοποίηση και το περιβάλλον χρήσης του διεγερόμενου από τη θεμελιώδη συχνότητα LPC (Pitch Excited LPC), καθώς και ορισμένα αποτελέσματα από τη χρήση αυτού του κωδικοποιητή.

ΚΕΦΑΛΑΙΟ 1

1.1 Η Ομιλία ως Μέσο Επικοινωνίας

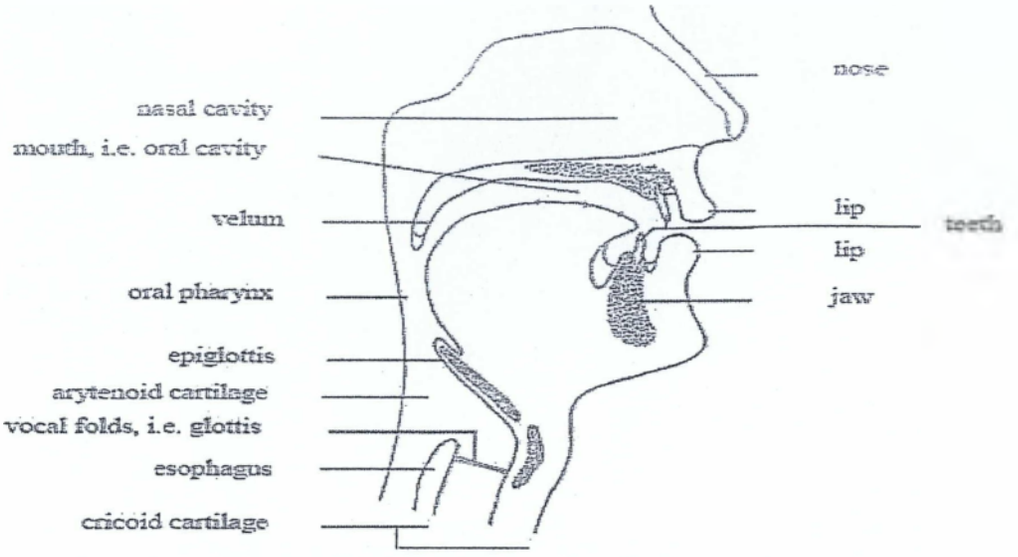
Στην πιο γενική μορφή της, η επικοινωνία με ομιλία είναι η διαδικασία της προφορικής μετάδοσης μιας ιδέας από έναν άνθρωπο σε έναν άλλο. Η διαδικασία υλοποίησης αυτής της επικοινωνίας αρχίζει με τη μετατροπή της ιδέας σε λογική πρόταση, που στη συνέχεια μετατρέπεται σε μυϊκές εκφράσεις της φωνητικής περιοχής, του λάρυγγα και των πνευμόνων. Η φωνητική περιοχή μετατρέπει την πρόταση σε ακουστικό κύμα πίεσης αέρα, το οποίο και γίνεται αντιληπτό από το ακουστικό σύστημα του ακροατή ως ομιλία. Με τη χρήση γνωστικών πηγών και τη γνώση της γλώσσας ο ακροατής κατανοεί την πρόταση και εξάγει το νόημα της.

Σχήμα 1.1: Διάγραμμα Μηχανισμού Ομιλίας ως Μέσο Επικοινωνίας.

Στη σημερινή εποχή όμως, η ανάγκη για επικοινωνία πέραν του κοντινού μας περιβάλλοντος και η ανάπτυξη της τεχνολογίας των τηλεπικοινωνιών έδωσαν μεγάλη ώθηση στην επιστήμη της ψηφιακής επεξεργασίας σήματος ομιλίας, δίνοντας μας την δυνατότητα να μεταφέρουμε τις ιδέες μας φωνητικά, οπουδήποτε επιθυμούμε.

1.2 Ανθρωπινός Μηχανισμός Παραγωγής Ομιλίας

Ανεξαρτήτως της ομιλούμενης γλώσσας όλοι οι άνθρωποι χρησιμοποιούν την ίδια ανατομία για την παραγωγή φωνής. Η παραγωγή της ομιλίας μπορεί να θεωρηθεί ότι αρχίζει στους πνεύμονες από όπου ξεκινάει και η ροή του αέρα. Ο πραγματικός όμως ήχος σχηματίζεται καθώς ο αέρας περνάει μέσα από το λάρυγγα και τη φωνητική περιοχή. Στο λάρυγγα βρίσκονται οι φωνητικές χορδές, ενώ η φωνητική περιοχή μπορεί να διαχωριστεί σε τρεις περιοχές: στο φάρυγγα, στη ρινική κοιλότητα και στο στόμα. Η γλώσσα, η υπερώα, η κάτω γνάθος και τα δόντια έχουν τη μεγαλύτερη επίδραση στη μορφή της φωνητικής περιοχής και έτσι στη διακριτότητα των ήχων.

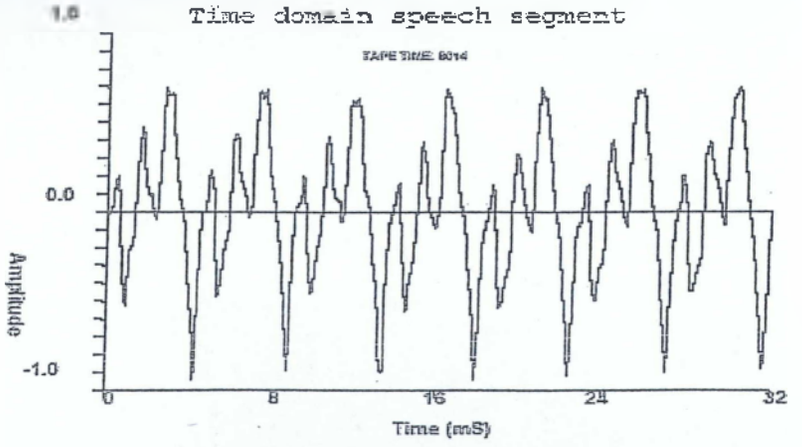


Σχήμα 1.2: Φωνητικός Μηχανισμός Ανθρώπου.

1.2.1 Ιδιότητες του Σήματος Ομιλίας

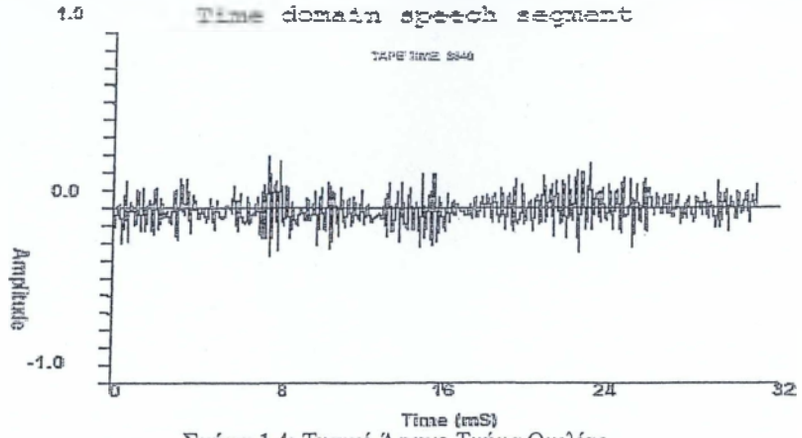
Τα σήματα ομιλίας δεν είναι στατικά και στην καλύτερη περίπτωση μπορούν να θεωρηθούν στατικά μόνο για πολύ μικρά διαστήματα της τάξης των 5-20 ms. Έτσι οι στατιστικές και φασματικές ιδιότητες της ομιλίας μπορούν να καθοριστούν σε αυτά τα μικρά διαστήματα. Η ομιλία μπορεί γενικά να κατηγοριοποιηθεί ως έμφωνη (voiced), άφωνη (unvoiced) ή μικτή.

Έμφωνη: Αυτή είναι συνήθως φωνήεντα (π.χ. [a], [e]) που παράγονται κατά το περιοδικό ανοιγοκλείμα των φωνητικών χορδών, δημιουργώντας έτσι μια σειρά από παλμούς (γλωττιδικός παλμός). Χαρακτηριστικά της έμφωνης ομιλίας είναι η περιοδικότητα και η αρμονική δομή στη συχνότητα.



Σχήμα 1.3: Τυπικό Έμφωνο Τμήμα Ομιλίας.

Άφωνη: Αυτή είναι συνήθως σύμφωνα (π.χ. [l], [r]) που παράγονται όταν οι φωνητικές παραμένουν ανοιχτές και ο αέρας περνάει ελεύθερα ανάμεσα τους δημιουργώντας ένα στροβιλισμό ανάμεσα στις χορδές και ο ήχος που προκύπτει μοιάζει με θόρυβο.



Σχήμα 1.4: Τυπικό Άφωνο Τμήμα Ομιλίας

Επιπλέον το ενεργειακό επίπεδο των έμφωνων τμημάτων είναι γενικά υψηλότερο από το ενεργειακό επίπεδο των άφωνων τμημάτων. Μια σημαντική ιδιότητα στο φάσμα της ομιλίας είναι οι περιοχές των συχνοτήτων που υπάρχει συγκεντρωμένη ενέργεια και ονομάζονται formants. Αυτά τα formants είναι γνωστό ότι παίζουν σημαντικό ρόλο στην αναγνώριση των ήχων της ομιλίας και κυρίως των φωνηέντων.

Δύο τύποι συσχέτισης παρουσιάζονται σε ένα σήμα. Αυτοί είναι γνωστοί ως πλεονασμοί κοντινών δειγμάτων και πλεονασμοί μακρινών δειγμάτων. Οι πλεονασμοί κοντινών δειγμάτων είναι αυτοί που είναι παρόντες ανάμεσα στα δείγματα ομιλίας που είναι πολύ κοντά μεταξύ τους. Οι πλεονασμοί των μακρινών δειγμάτων προέρχονται από την κληρονομούμενη περιοδικότητα της έμφωνης ομιλίας.

Αυτές οι ιδιότητες της ομιλίας είναι βασικές για την κωδικοποίηση και τη συμπίεση της ομιλίας.

1.3 Στόχος των Κωδικοποιητών Φωνής

Ο στόχος των κωδικοποιητών φωνής είναι να αναπαραστήσουν την ομιλία σε ψηφιακή μορφή με την καλύτερη δυνατή ποιότητα και με το μικρότερο αριθμό bits. Οι περισσότεροι κωδικοποιητές ομιλίας βασίζονται σε αλγόριθμους συμπίεσης με απώλειες όπου το σημασιολογικό περιεχόμενο δεν αλλοιώνεται αλλά η ποιότητα της ομιλίας μειώνεται. Αυτοί οι κωδικοποιητές είναι αποδεκτοί λόγω του ότι η μείωση της ποιότητας συχνά δεν γίνεται αντιληπτή από το ανθρώπινο ακουστικό σύστημα. Ο κατώτερος αριθμός bits με τον οποίο μπορεί να κωδικοποιηθεί το σήμα της ομιλίας καθορίζεται από τη φωνημική πληροφορία η οποία έχει εκτιμηθεί ότι είναι περίπου 50 bits ανά δευτερόλεπτο, και από το συνολικό ρυθμό γνωστικής πληροφορίας σε ένα σήμα ομιλίας και έχει εκτιμηθεί ότι είναι περίπου 400 bits ανά δευτερόλεπτο.

Ακόμα και αν υπήρχε αυτός ο ιδανικός κωδικοποιητής ομιλίας με ιδανική ποιότητα στα 400 bits ανά δευτερόλεπτο θα ήταν δύσκολο να επιλεχθεί σε σχέση με κάποιον άλλο όχι και τόσο καλό κωδικοποιητή ομιλίας. Αυτό οφείλεται στο ότι αυτοί οι

κωδικοποιητές είναι ιδιαίτερα ευαίσθητοι σε μη ακουστικά σήματα ομιλίας και σε σήματα ομιλίας με θόρυβο. Επίσης δεν είναι ιδιαίτερα ανθεκτικοί σε σφάλματα καναλιού και μπορούν να παρουσιάσουν μεγάλες καθυστερήσεις κατά την επεξεργασία του σήματος ομιλίας.

Η επιλογή της χρήσης ενός κωδικοποιητή γίνεται ανάλογο με τις συνθήκες στις οποίες αυτός θα χρησιμοποιηθεί είτε αυτή είναι μετάδοση είτε αυτή είναι αποθήκευση. Για την περίπτωση της μετάδοσης μας ενδιαφέρει ο κωδικοποιητής να εισάγει όσο το δυνατό λιγότερη καθυστέρηση ιδικά όταν σε αυτή υπάρχουν διάφορες επιπρόσθετες καθυστερήσεις, ενώ για την αποθήκευση η παράμετρος της καθυστέρησης δενν είναι μεγάλης σημασίας. Εδώ μας ενδιαφέρει ο κωδικοποιητής να έχει την ικανότητα της πρόγνωσης και της διόρθωσης των σφαλμάτων που μπορεί να παρουσιαστούν.

1.4 Μοντέλα Κωδικοποίησης Φωνής

Η κωδικοποίηση φωνής βασίζεται σε ορισμένα βασικά μοντέλα παραγωγής της ομιλίας τα οποία κατατάσσονται ως εξής:

Γνωστικό Μοντέλο: Αυτό βασίζεται στις διεργασίες που λαμβάνουν χώρα στον εγκέφαλο για την παραγωγή της ομιλίας. Το μοντέλο αυτό βρίσκεται ακόμα στο στάδιο της έρευνας και δεν έχει ακόμα εφαρμοσθεί σε ένα σύστημα κωδικοποιητή ομιλίας.

Γλωσσικό Μοντέλο: Το μοντέλο αυτό βασίζεται στο ότι ο προφορικός λόγος παρουσιάζει ακριβή και ολοκληρωμένη συντακτική δομή η οποία σχετίζεται με ένα μοτίβο συλλαβικής έντασης. Σε αυτόν δεν υπάρχει μόνο η πληροφορία για το νοηματικό περιεχόμενο αλλά και η στάση του ομιλητή προς αυτό. Οι κωδικοποιητές που κάνουν χρήση αυτού του μοντέλου παράγουν ομιλία χαμηλής ποιότητας αλλά μας δίνουν την δυνατότητα να απομονώσουμε στοιχεία της ομιλίας (π.χ. το νόημα της πρότασης, τη στάση του ομιλητή) και έτσι να μπορέσουμε να εστιάσουμε αποκλειστικά στα χαρακτηριστικά της κυματομορφής της ομιλίας που μας ενδιαφέρουν.

Μοντέλο Φωνητικού Σωλήνα: Το μοντέλο αυτό βασίζεται στον τρόπο με τον οποίο λειτουργεί η φωνητική περιοχή του ανθρώπου και στους βασικούς ήχους που αυτή μπορεί να παράγει, έμφωνα – άφωνα. Οι κωδικοποιητές που χρησιμοποιούν το μοντέλο αυτό προσπαθούν να μοντελοποιήσουν τη διεργασία παραγωγής της φωνής με ένα δυναμικό σύστημα και επιπλέον προσπαθούν να ποσοτικοποιήσουν συγκεκριμένους περιορισμούς γι' αυτό το σύστημα. Βασικές λειτουργίες αυτών των κωδικοποιητών είναι να αναλύουν το σήμα της ομιλία στον πομπό, να μεταδίδουν τις παραμέτρους που προκύπτουν από την ανάλυση και έπειτα χρησιμοποιώντας τις παραμέτρους αυτές να επανασυνθέτουν την ομιλία στο δέκτη. Οι κωδικοποιητές αυτοί αποδίδουν μικρό αριθμό bits με όχι απαραίτητα καλή ποιότητα. Το μοντέλο παρουσιάζεται πιο αναλυτικά στο επόμενο κεφάλαιο.

Ακουστικό Μοντέλο: Μια κοινή χαρακτηριστική ιδιότητα όλων των κωδικοποιητών είναι ότι παράγουν σήματα που θα ληφθούν από το ανθρώπινο ακουστικό σύστημα. Έτσι λοιπόν αν οποιαδήποτε πληροφορία στο αυθεντικό σήμα ομιλίας φιλτραριστεί και απορριφθεί από το ανθρώπινο ακουστικό σύστημα , τότε αυτή η πληροφορία μπορεί να αφεθεί εκτός της κωδικοποιημένης αναπαράστασης του

σήματος, με αποτέλεσμα τα bits που είναι απαραίτητα για να αναπαρασταθεί το σήμα να μειωθούν.

Ένα μοντέλο που έχει χρησιμοποιηθεί επιτυχώς από τους κωδικοποιητές ομιλίας είναι το κρίσιμης ζώνης (critical band) μοντέλο για την ακουστική αντίληψη. Το μοντέλο αυτό προσπαθεί να συγκρατήσει έναν αριθμό συσχετιζόμενων στοιχείων της ακουστικής αντίληψης. Το πρώτο στοιχείο είναι η ανάλυση ακουστικής συχνότητας. Το εύρος μιας κρίσιμης ζώνης σε μια συγκεκριμένη συχνότητα είναι το μέτρο του πόσο απομακρυσμένοι δύο τόνοι πρέπει να είναι στη συχνότητα, ώστε αυτοί οι δύο τόνοι να είναι διακριτοί μεταξύ τους. Το δεύτερο στοιχείο είναι η ακουστική επικάλυψη θορύβου. Αυτό το στοιχείο μπορεί να διατυπωθεί ως εξής: Ένα σήμα ομιλίας σε μια συγκεκριμένη κρίσιμη ζώνη θα επικαλύψει ένα σήμα (θόρυβος) που βρίσκεται στην ίδια ζώνη. Έτσι σήματα θορύβου που βρίσκονται κοντά στο σήμα ομιλίας στο πεδίο της συχνότητας καλύπτονται.

1.5 Καθορισμός Ενός Κωδικοποιητή Φωνής

Ένας κωδικοποιητής ομιλίας γενικά αποτελείται από τρία στάδια: την ανάλυση ομιλίας, την κβάντιση των παραμέτρων και την κωδικοποίηση των παραμέτρων. Η έξοδος του πρώτου σταδίου εξαρτάται από τον κωδικοποιητή που χρησιμοποιούμε και τον τρόπο με τον οποίο αυτός μοντελοποιεί το σήμα της ομιλίας. Έτσι για παράδειγμα, σε ένα PCM σύστημα το σήμα της ομιλίας δεν υπόκειται σε καμία ανάλυση καθώς η έξοδος του, είναι απλά το ψηφιακό σήμα της ομιλίας. Ενώ για ένα LPC σύστημα η έξοδος του πρώτου σταδίου θα είναι οι παράμετροι του σήματος ομιλίας.

Μετά την ανάλυση οι παράμετροι πρέπει να κβαντιστούν για να μειωθεί ο αριθμός των bits που απαιτούνται για την αναπαράσταση του σήματος της ομιλίας. Η έξοδος του σταδίου κβάντισης μπορεί να θεωρηθεί ως μια θορυβημένη αναπαράσταση της εισόδου και είναι μια μη αντιστρεπτή διαδικασία. Το τρίτο στάδιο είναι η κωδικοποίηση του κβαντισμένου σήματος κατά την οποίας ο κωδικοποιητής δίνει μια μοναδική δυαδική τιμή σε κάθε πιθανή κβαντισμένη αναπαράσταση. Συνήθως αυτοί οι δυαδικοί αριθμοί συνδυάζονται σε πακέτα για την αποτελεσματικότερη αποθήκευση και μετάδοση.

Ο αποκωδικοποιητής ομιλίας αντιστρέφει τις διεργασίες του κωδικοποιητή. Αφού το κωδικοποιημένο σήμα αποκωδικοποιηθεί θα εξαχθούν από αυτό οι παράμετροι του σήματος ομιλίας μέσω ενός αντίστροφου κβαντιστή. Οι παράμετροι αυτοί, απουσία bit error, θα συντεθούν και θα μας δώσουν το αρχικό μας σήμα.

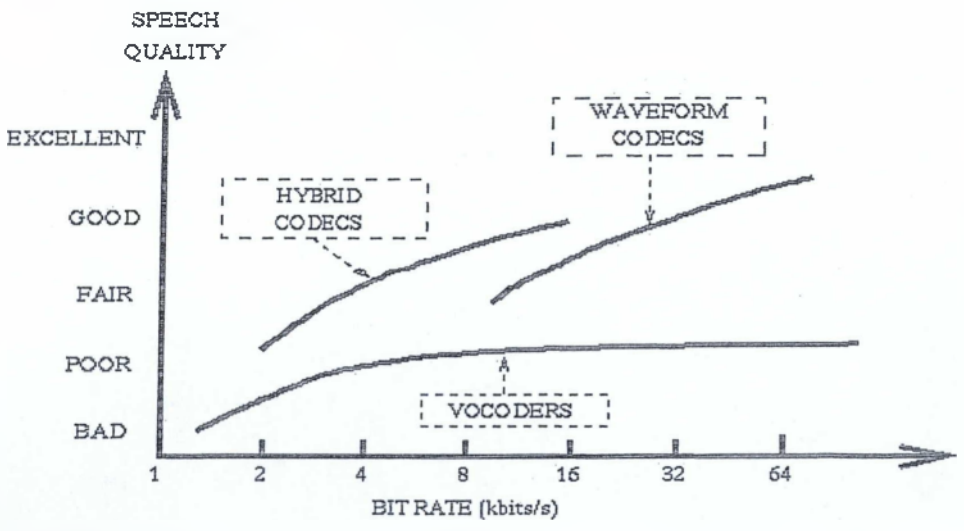
1.6 Κατηγορίες Κωδικοποιητών Φωνής

Ανάλογα με τα χαρακτηριστικά του σήματος ομιλίας (Παράγραφος 1.2.1) που χρησιμοποιούν οι κωδικοποιητές ομιλίας, καθώς και με το αν κάνουν χρήση κάποιου μοντέλου κωδικοποίησης φωνής (Παράγραφος 1.4) οι κωδικοποιητές ομιλίας μπορούν να ταξινομηθούν σε τρεις γενικές κατηγορίες:

Κωδικοποιητές Κυματομορφής: Οι κωδικοποιητές αυτοί δεν απευθύνονται αποκλειστικά στην κωδικοποίηση ομιλίας. Σκοπός αυτής της κατηγορίας κωδικοποιητών είναι να αναπαραστήσουν το αυθεντικό σήμα ομιλίας (ή γενικά οποιοδήποτε άλλο σήμα) με τη μεγαλύτερη δυνατή ακρίβεια.

Vocoder Κωδικοποιητές: Η κατηγορία αυτή των κωδικοποιητών αφορά αποκλειστικά την κωδικοποίηση φωνής και βασίζεται στην παραμετρική περιγραφή της ομιλίας. Σκοπός αυτής της κατηγορίας είναι να παράγει αντιληπτά κατανοήσιμη ομιλία χωρίς απαραίτητα να διατηρεί την κυματομορφή του αυθεντικού σήματος. Κατά κανόνα το παραγόμενο σήμα των κωδικοποιητών αυτών έχει αφύσικη ή συνθετική χροιά, ενώ επιτυγχάνεται χαμηλότερο bit rate σε σχέση με τους κωδικοποιητές κυματομορφής.

Υβριδικοί Κωδικοποιητές: Η κατηγορία αυτή των κωδικοποιητών συνδυάζει τα χαρακτηριστικά των δύο παραπάνω κατηγοριών. Έτσι έχουν και την ικανότητα της αποτελεσματικής κωδικοποίησης των vocoders αλλά και την ποιότητα σήματος των κωδικοποιητών κυματομορφής.



Σχήμα 1.5: Ταξινόμηση των κωδικοποιητών σε σχέση με το bit rate και την ποιότητα ομιλίας.

Τις τρεις αυτές σημαντικές κατηγορίες κωδικοποιητών θα τις δούμε λίγο πιο αναλυτικά στις επόμενες παραγράφους.

1.6.1 Κωδικοποιητές Κυματομορφής

Καθώς ο απόλυτος στόχος των κωδικοποιητών κυματομορφής είναι να αναπαραστήσουν όσο το δυνατόν πιστότερα το αυθεντικό σήμα ομιλίας δείγμα προς δείγμα, η κατηγορία αυτών είναι ιδιαίτερα ανθεκτική σε διάφορους τύπους εισόδων. Ο πιο απλός κωδικοποιητής της κατηγορίας αυτής είναι η Παλμοκωδική διαμόρφωση (PCM) που χρησιμοποιεί ένα σταθερό κβαντιστή για κάθε δείγμα του σήματος ομιλίας. Δεδομένης της μη-ομοιόμορφης κατανομής των πλατών στο δείγμα ομιλίας και της λογαριθμικής ευαισθησίας του ανθρώπινου ακουστικού συστήματος, ένας μη-ομοιόμορφος κβαντιστής μπορεί να μας δώσει καλύτερη ποιότητα από έναν ομοιόμορφο κβαντιστή με το ίδιο bit rate. Έτσι έχουμε τον λογαριθμικό PCM με toll ποιότητα ομιλίας.

Η toll ποιότητα ομιλίας μπορεί να επιτευχθεί με αρκετά χαμηλότερο bit rate έχοντας ως κόστος όμως μεγαλύτερη πολυπλοκότητα. Ένας τέτοιος κωδικοποιητής είναι ο Προσαρμοστικός Διαφορικός PCM (ADPCM). Σε αυτό το δείγμα ομιλίας έχει

προβλεφθεί από προηγούμενα δείγματα και το σφάλμα της πρόγνωσης κβαντίζεται. Και ο προγνώστης και ο κβαντιστής μπορούν να προσαρμοσθούν για να βελτιωθεί η απόδοση. Μια άλλη πιθανότητα είναι να μετατρέψουμε το σήμα ομιλίας σε ένα άλλο πεδίο με ένα διακριτό συνημιτονοειδή μετασχηματισμό (DCT) ή με έναν άλλο κατάλληλο μετασχηματισμό. Ο μετασχηματισμός συμπυκνώνει την ενέργεια σε λίγους συντελεστές που μπορούν να κβαντιστούν αποτελεσματικά. Στην κατηγορία αυτή εμπίπτει ο κωδικοποιητής προσαρμοστικού μετασχηματισμού (ATC). Εδώ ο κβαντιστής μπορεί να προσαρμοσθεί ανάλογα με τα χαρακτηριστικά του σήματος.

1.6.2 Vocoder Κωδικοποιητές

Η αποτελεσματικότητα των vocoders είναι ισχυρά εξαρτώμενη από την ακρίβεια των μοντέλων κωδικοποίησης φωνής. Οι κωδικοποιητές αυτοί έχουν σχεδιαστεί για εφαρμογές χαμηλού bit rate (όπως στρατιωτικές ή δορυφορικές επικοινωνίες) και ο βασικός τους στόχος είναι να διατηρούν την κατανοησιμότητα της ομιλίας. Οι περισσότερο αποτελεσματικοί vocoders βασίζονται στην κωδικοποίηση γραμμικής πρόγνωσης (LPC). Με τον LPC, κάθε πλαίσιο ομιλίας μοντελοποιείται ως η έξοδος ενός γραμμικού συστήματος που αναπαριστά τη φωνητική περιοχή, σε ένα σήμα διέγερσης. Οι παράμετροι αυτού του συστήματος και η διέγερση του κωδικοποιούνται και μεταδίδονται. Vocoders βασισμένοι στον LPC μπορούν να επιτύχουν communication ποιότητα ομιλίας με ρυθμούς κάτω από 2 kb/s.

1.6.3 Υβριδικοί Κωδικοποιητές

Ενώ η ποιότητα των κωδικοποιητών κυματομορφής πέφτει ραγδαία για ρυθμούς μετάδοσης κάτω των 16 kb/s, η βελτίωση της ποιότητας των vocoder είναι αμελητέα για ρυθμούς πάνω από 4 kb/s. Εδώ έρχονται οι υβριδικοί κωδικοποιητές για να γεφυρώσουν αυτό το χάσμα, παρέχοντας καλή ποιότητα ομιλίας με μέσους ρυθμούς μετάδοσης, παρουσιάζοντας όμως υψηλότερες υπολογιστικές απαιτήσεις.

Οι υβριδικοί κωδικοποιητές ανήκουν στην κατηγορία των analysis-by-synthesis, χρησιμοποιώντας την LPC ανάλυση για να πάρουν τις παραμέτρους του μοντέλου σύνθεσης. Στη συνέχεια εφαρμόζονται τεχνικές κωδικοποίησης κυματομορφής για την κωδικοποίηση του σήματος διέγερσης. Κωδικοποιητές αυτής της κατηγορίας είναι ο Πολλαπλών Παλμών Διέγερσης (MPE), ο Τακτικού Παλμού Διέγερσης (RPE) και ο κωδικοποιητής Γραμμικής Πρόγνωσης με τη χρήση Codebook Διέγερσης (CELP).

Οι vocoders και οι υβριδικοί κωδικοποιητές της κατηγορίας analysis-by-synthesis θα αναλυθούν περαιτέρω στα επόμενα κεφάλαια.

1.7 Παράμετροι της Απόδοσης των Κωδικοποιητών Φωνής

α) Ρυθμός μετάδοσης: Η ψηφιακή κωδικοποίηση χαρακτηρίζεται από απώλεια πληροφορίας (lossy compression). Στόχος είναι η επίτευξη της ελάχιστης απώλειας για ένα δοσμένο ρυθμό μετάδοσης.

β) Ποιότητα φωνής: Εκτιμάται με βάση τη φυσικότητα και την ευκολία κατανόησης (intelligibility) της κωδικοποιημένης φωνής. Υπάρχουν αντικειμενικά και υποκειμενικά μέτρα παραμόρφωσης.

Αντικειμενικά μέτρα παραμόρφωσης (Objective distortion measures): Αυτά πρέπει:

1. Να αντιστοιχούν σε ψυχολογικά χαρακτηριστικά της ανθρώπινης αντίληψης
2. Να υπολογίζονται εύκολα και αποτελεσματικά
3. $d(x1, x2) \geq 0$

Υποκειμενικά μέτρα παραμόρφωσης (Subjective distortion measures):

Mean-opinion score (MOS): Η ποιότητα φωνής αξιολογείται συνήθως σε μια κλίμακα πέντε επιπέδων, γνωστή ως κλίμακα MOS, η οποία προέρχεται από τον μέσο όρο των δεδομένων φωνής, των ομιλητών και των ακροατών. Τα πέντε επίπεδα της ποιότητας είναι: Κακό (bad), ανεπαρκές (poor), μέτριο (fair), καλό (good) και άριστο (excellent). Η ποιότητα θεωρείται ικανοποιητική για βαθμό 3,5 ή υψηλότερο που γενικά υπονοεί υψηλά επίπεδα σαφήνειας, αναγνώρισης ομιλητών και φυσικότητας (Βαθμολογία: 1-Bad, 2-Poor, 3-Fair, 4-Good, 5-Excellent).

γ) Καθυστέρηση κωδικοποιητή (Encoding Delay):

Δημιουργείται από την πολυπλοκότητα του κωδικοποιητή και προστίθεται στις καθυστερήσεις μετάδοσης (π.χ. δορυφορικά κανάλια) και την κωδικοποίηση πηγής. Προξενεί σημαντικά προβλήματα στην επικοινωνία καθώς:

1. Μπορεί να δημιουργηθεί ηχώ (σε μετασχηματιστές δισύρματων σε τετρασυρμάτινες ζεύξεις)
2. Παρεμποδίζει την επικοινωνία

Επιτρεπτές τιμές: $\leq$ 1ms στην τηλεφωνία, $\leq$ 500ms video telephony, αυθαίρετη για αποθήκευση (π.χ. για voice mail)

δ) Ευαισθησία στα σφάλματα μετάδοσης:

1. Μπορεί να έχουμε bursts (mobile communications)
2. Οι κωδικοποιητές αφαιρούν περιττή πληροφορία σε μικρότερους ρυθμούς μετάδοσης και συνήθως αυξάνεται η ευαισθησία στα σφάλματα μετάδ.
3. Υποκειμενική εκτίμηση των σφαλμάτων μετάδοσης.

ε) Πολυπλοκότητα:

Η πολυπλοκότητα ενός αλγορίθμου κωδικοποίησης είναι η προσπάθεια επεξεργασίας που απαιτείται για να αναπαρασταθεί ο αλγόριθμος και τυπικά μετριέται σε σχέση με την αριθμητική δυνατότητα και τις απαιτήσεις μνήμης ή ισοδύναμα από την άποψη του κόστους. Μια μεγάλη πολυπλοκότητα μπορεί να οδηγήσει στη μεγάλη κατανάλωση ενέργειας στο hardware.

1.8 Αξιολόγηση της Απόδοσης Κωδικοποιητών Φωνής

Για να αξιολογήσουμε την απόδοση ενός κωδικοποιητή φωνής είναι απαραίτητο να έχουμε κάποιο δείκτη της κατανοησιμότητας και της ποιότητας ομιλίας που παράγεται. Ο όρος κατανοησιμότητας αναφέρεται στο αν η παραγόμενη ομιλία είναι εύκολα κατανοητή ενώ η ποιότητα είναι ένας δείκτης που δείχνει πόσο φυσικά ακούγεται η ομιλία.

Οι τεχνικές αξιολόγησης της απόδοσης ενός κωδικοποιητή φωνής όσον αφορά την κατανοησιμότητα και την ποιότητα της ομιλίας που αυτός παράγει χωρίζονται σε δύο

κατηγορίες τις αντικειμενικές και τις υποκειμενικές.

1.8.1 Αντικειμενικές Τεχνικές Αξιολόγησης Κωδικοποιητών Φωνής

Σε αυτές τις τεχνικές γίνεται σύγκριση μεταξύ, του αυθεντικού σήματος ομιλίας και του κωδικοποιημένου σήματος ομιλίας. Η πιο συνηθισμένη τεχνική είναι η τεχνική του λόγου σήματος προς θόρυβο (Signal to Noise Ratio) που δίνεται από τη σχέση

$$SNR = 10\log_{10}\left\{\frac{\sum_{n=0}^{M} s^2(n)}{\sum_{n=0}^{M}\left(s(n)-\bar{s}(n)\right)^2}\right\} \quad (1.1)$$

όπου το $s(n)$ είναι το αυθεντικό σήμα ομιλίας και $\bar{s}(n)$ είναι το κωδικοποιημένο σήμα. Η τεχνική SNR χρησιμοποιείται για μακροπρόθεσμες μετρήσεις της απόδοσης του συστήματος και για αυτό είναι απαραίτητη η χρήση μιας τεχνικής που βασίζεται στο βραχυπρόθεσμο λόγο σήματος προς θόρυβο. Αυτή η τεχνική απόδοσης είναι το τμηματικό SNR (Segmental Signal to Noise Ratio) ή SEGSNR και δίνεται από τη σχέση

$$SEGSNR = \frac{10}{L}\sum_{i=0}^{L-1}\log_{10}\left\{\frac{\sum_{n=0}^{M} s^2(iN+n)}{\sum_{n=0}^{M}\left(s(iN+n)-\bar{s}(iN+n)\right)^2}\right\} \quad (1.2)$$

1.8.2 Υποκειμενικές Τεχνικές Αξιολόγησης Κωδικοποιητών Φωνής

Επειδή οι αντικειμενικές μετρήσεις είναι ευαίσθητες στις μεταβολές κέρδους και στις καθυστερήσεις, και κυρίως επειδή δεν παίρνουν υπόψη τους τις perceptual ιδιότητες του αυτιού, είναι απαραίτητο να χρησιμοποιήσουμε υποκειμενικές τεχνικές αξιολόγησης. Αυτές οι τεχνικές βασίζονται αποκλειστικά στον ανθρώπινο παράγονται και γίνονται με την παρουσίαση μιας φωνητικά ισορροπημένης ομιλίας σε ένα ακροατήριο που καλείται να κρίνει αυτή την ομιλία. Οι πιο διαδεδομένες τεχνικές υποκειμενικής αξιολόγησης είναι οι εξής:

Diagnostic Rhyme Test (DRT): Το διαγνωστικό τεστ ρίμας μετράει την καταληπτότητα. Σε αυτή την τεχνική το ακροατήριο καλείται να αναγνωρίσει μία από τις δύο πιθανές λέξεις όταν του παρουσιάζεται ένα ζεύγος λέξεων που κάνουν ρίμα όπως

π-ονος μ - ονος
π-ερας, τ-ερας

Το τελικό αποτέλεσμα του διαγνωστικού τεστ ρίμας είναι επί της εκατό και δίνεται από τη σχέση

$$P = \frac{(R - W) \times 100}{T} \qquad (1.3)$$

όπου R είναι των λέξεων που επιλέχθηκαν σωστά, W είναι α αριθμός των λέξεων που επιλέχθηκαν λάθος και T είναι ο συνολικός αριθμός των ζευγών λέξεων που ελέχθησαν. Το συγκεκριμένο τεστ διενεργείται κυρίως σε κωδικοποιητές που παράγουν χαμηλής ποιότητας ομιλία και συνήθως ισχύει $75 \leq DRT \leq 95$ με το 90 να ανταποκρίνεται σε ένα καλό σύστημα.

Diagnostic Acceptability Measure (DAM): Στην τεχνική του μέτρου διαγνωστικής αποδεκτικότητας το ακροατήριο καλείται να αξιολογήσει ισορροπημένες προτάσεις από τον κωδικοποιητή ενδιαφέροντος προσδίδοντας σε αυτές έναν αριθμό μεταξύ του 0 και 100 σε τρεις κατηγορίες: ποιότητα σήματος, ποιότητα υποβάθρου και ολική επίδραση. Οι εκτιμήσεις κάθε κατηγορίες σταθμίζονται κατάλληλα και μας δίνουν το αποτέλεσμα. Ένα τυπικό DAM αποτέλεσμα είναι 45 – 55% με το 50% να ανταποκρίνεται σε ένα καλό σύστημα.

Mean Opinion Score (MOS): Στην τεχνική MOS το ακροατήριο καλείται να βαθμολογήσει την κωδικοποιημένη ομιλία σε μια κλίμακα από το 1 έως το 5, κατατάσσοντας την έτσι σε μια από τις παρακάτω πέντε αντίστοιχες κατηγορίες κατηγορίες: κακή, φτωχή, μέτρια, καλή και εξαιρετική όπως φαίνεται και από τον Πίνακα 1.1.

Κλίμακα MOS	Ποιότητα Ομιλίας
1	Κακή
2	Φτωχή
3	Μέτρια
4	Καλή
5	Εξαιρετική

Πίνακας 1.1: Κλίμακα MOS

Στο τέλος υπολογίζεται ο μέσος όρος από τη βαθμολογία όλων των ακροατών για να πάρουμε την αξιολόγηση MOS για τον εκάστοτε κωδικοποιητή. Μια αξιολόγηση MOS με τιμή 4 έως 4.5 υποδηλώνει υψηλή ποιότητα.

1.9 Κατηγορίες Ποιότητας της Ομιλίας

Στις ψηφιακές επικοινωνίες η ποιότητα της ομιλίας ταξινομείται σε τέσσερις διαφορετικές κατηγορίες, αυτές είναι:

Broadcast: Η Broadcast ευρείας ζώνης ομιλία αναφέρεται σε υψηλής ποιότητα

ομιλία που γενικά μπορεί να επιτευχθεί με ρυθμούς μετάδοσης άνω των 64 kb/s.

Toll ή Network: Αυτή αναφέρεται σε ποιότητα που μπορεί να συγκριθεί με την κλασσική αναλογική ομιλία (200-300 Hz) και επιτυγχάνεται με ρυθμούς μετάδοσης άνω των 16 kb/s.

Communication: Αυτή υπονοεί κάποια υποβάθμιση στην ποιότητα ομιλίας, η οποία όμως παραμένει φυσική, υψηλά αναγνωρίσιμη, και επαρκής για τηλεπικοινωνίες. Αυτή η ποιότητα επιτυγχάνεται με ρυθμούς μετάδοσης πάνω από 4.8 kb/s.

Synthetic: Αυτή η ποιότητα μας δίνει συνθετική ομιλία η οποία είναι μεν κατανοητή αλλά ο ομιλητής τείνει να μην αναγνωρίζεται.

ομιλία που γενικά μπορεί να επιτευχθεί με ρυθμούς μετάδοσης άνω των 64 kb/s.

Toll ή Network: Αυτή αναφέρεται σε ποιότητα που μπορεί να συγκριθεί με την κλασσική αναλογική ομιλία (200-300 Hz) και επιτυγχάνεται με ρυθμούς μετάδοσης άνω των 16 kb/s.

Communication: Αυτή υπονοεί κάποια υποβάθμιση στην ποιότητα ομιλίας, η οποία όμως παραμένει φυσική, υψηλά αναγνωρίσιμη, και επαρκής για τηλεπικοινωνίες. Αυτή η ποιότητα επιτυγχάνεται με ρυθμούς μετάδοσης πάνω από 4.8 kb/s.

Synthetic: Αυτή η ποιότητα μας δίνει συνθετική ομιλία η οποία είναι μεν κατανοητή αλλά ο ομιλητής τείνει να μην αναγνωρίζεται.

ΚΕΦΑΛΑΙΟ 2

2.1 Εισαγωγή

Η γραμμικής πρόβλεψης κωδικοποίηση (LPC) είναι μια από τις πιο δημοφιλείς τεχνικές κωδικοποίησης για σήματα ομιλίας. Ο LPC δεν τόσο ένας τύπος κωδικοποιητή αλλά μια τεχνική που χρησιμοποιείται σε μια πληθώρα από διαφορετικούς τύπους κωδικοποιητών φωνής. Μπορεί να χρησιμοποιηθεί (και χρησιμοποιείται) στους διεγερόμενους θεμελιώδους συχνότητας vocoders, στους φωνητικά διεγερόμενους vocoders, στους κωδικοποιητές κυματομορφής, στους analysis-bysynthesis κωδικοποιητές, ακόμα και στους κωδικοποιητές του πεδίου συχνότητας. Οι κωδικοποιητές γραμμικής πρόβλεψης διεγερόμενης θεμελιώδους συχνότητας (pitch-exited linear predictive coders) έχουν το πλεονέκτημα ότι μπορούν να λειτουργούν με χαμηλούς ρυθμούς μετάδοσης, σχετικά μικρές υπολογιστικές πηγές και να παράγουν εύχρηστες κωδικοποιημένες αναπαραστάσεις του αυθεντικού σήματος ομιλίας. Το κύριο μειονέκτημα τους είναι ότι το μοντέλο διεγερόμενης θεμελιώδους συχνότητας περιορίζει τη μέγιστη ποιότητα του κωδικοποιητή ασχέτως των bit που θα χρησιμοποιηθούν.

2.2 Διεγερόμενος Θεμελιώδους Συχνότητας LPC (Pitch Excited LPC)

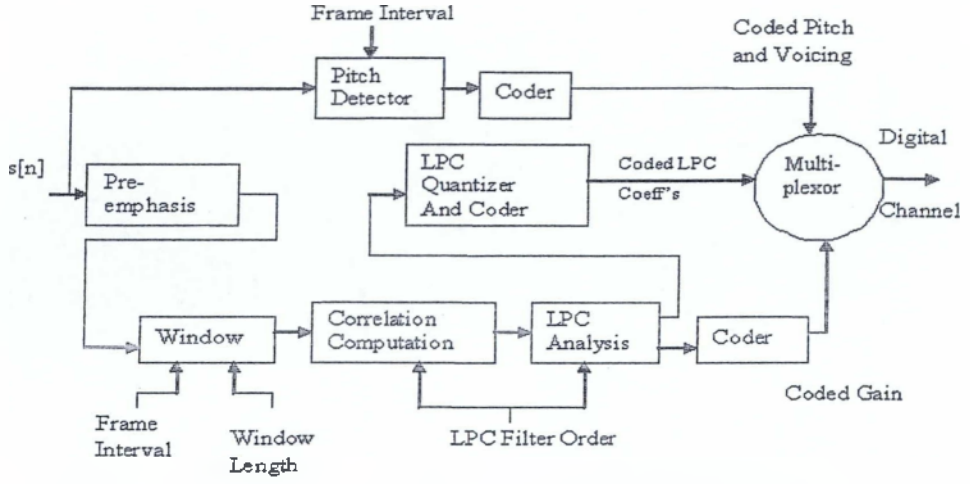
Όπως και όλοι οι άλλοι διεγερόμενοι από τη θεμελιώδη συχνότητα vocoders, έτσι και ο διεγερόμενος από τη θεμελιώδη συχνότητα LPC είναι ένας πλήρως παραμετροποιημένος κωδικοποιητής. Αυτό σημαίνει ότι η κωδικοποιημένη ομιλία χαρακτηρίζεται εξολοκλήρου από τις χρονικά μεταβαλλόμενες παραμέτρους ενός μοντέλου 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 διεγερόμενους 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 διεγερόμενου 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 διεγερόμενος 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.

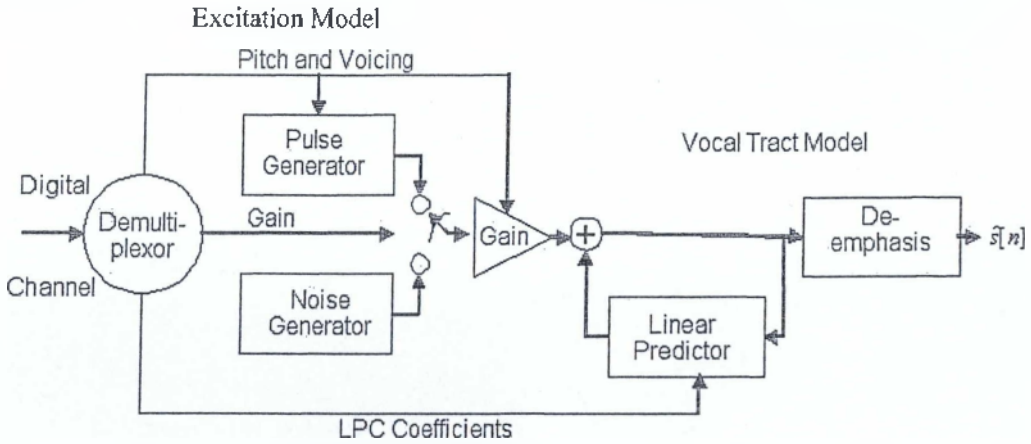


Σχήμα 2.1: Μπλοκ διάγραμμα ενός διεγερόμενου από τη θεμελιώδη συχνότητα LPC πομπού.

Το μπλοκ διάγραμμα ενός δέκτη γραμμικής πρόβλεψης διεγερόμενου 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 προβλέπτη 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 εμφώνων. 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.

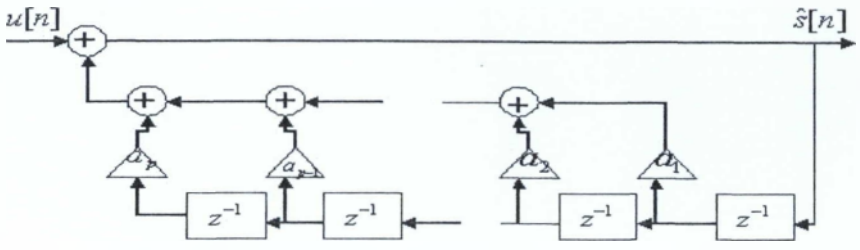


Σχήμα 2.2: Μπλοκ διάγραμμα ενός LPC δέκτη διεγερόμενου από τη θεμελιώδη συχνότητα.

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.

2.3 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.

2.3.1 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 αυτο-οπισθοδρομική (autoregressive) διαδικασία που μπορεί να αναπαρασταθεί ως

$$\hat{s}[n] = \sum_{i=1}^{P} a_i \hat{s}[n-i] + Gu[n] \quad \textbf{(2.1)}$$

όπου $\hat{s}[n]$ είναι η συνθετική ομιλία που παράγεται από το μοντέλο, $u[n]$ είναι το σήμα διέγερσης, a_i είναι οι παράμετροι πρόγνωσης και P η τάξη του προγνώστη.

Σε αυτή την έκφραση, G είναι η παράμετρος κέρδους που χρησιμοποιείται to match the energy of the synthetic speech to that of the original speech signal. Στο πεδίο των z-μετασχηματισμών, $\hat{S}(z)$ είναι η έξοδος του φίλτρου, $H(z)$, στο σήμα εισόδου, $U(z)$. Το LPC φίλτρο σύνθεσης $H(z)$ δίνεται από τη σχέση

$$\boldsymbol{H(z) = 1/(1 - A(z))} \quad \textbf{(2.2)}$$

Όπου A(z) είναι το φίλτρο του προγνώστη και δίνεται

$$\boldsymbol{A(z) = \sum_{i=1}^{P} a_k z^{-k}} \quad \textbf{(2.3)}$$

Με αυτούς τους όρους, το $\hat{S}(z)$ δίνεται από τον τύπο

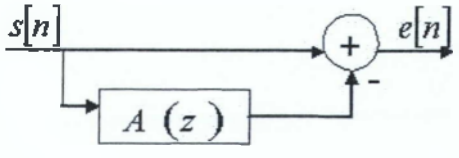
$$\hat{S}(z) = H(z) * U(z) = \frac{1}{1-A(z)} = \frac{U(z)}{1-\sum_{k=1}^{P} a_k z^{-k}} \qquad \textbf{(2.4)}$$

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 είναι οι συντελεστές του προγνώστη (ai ,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 στατικότητας μικρού χρόνου υπόθεση 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 στοπαράθυρο. Οι συντελεστές του A(z) , a_i , i = 1,..., 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, $\hat{s}(n)$, από το εισερχόμενο σήμα ομιλίας, $s(n)$. 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

$$\frac{1}{H(z)} = 1 - A(z). \qquad (2.5)$$

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

$$\mathbf{E} = \sum_n e^2[n], \quad (2.6)$$

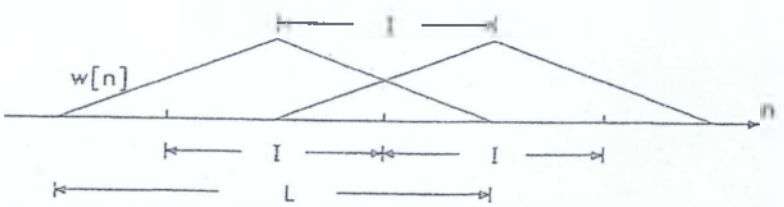
ως προς στους συντελεστές πρόγνωσης. Σε αυτή την έκφραση $e[n]$ είναι η έξοδος του αντίστροφου φίλτρου που δίνεται από

$$e[n] = s[n] - \sum_{i=1}^{p} a_i s[n-1], \qquad (2.7)$$

Υπάρχουν πολλοί μέθοδοι για να πάρουμε του 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

$$s_i[n] = s[n]w_i[n], \qquad (2.8)$$

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

$$w_i[n] = w[n - il], \quad (2.9)$$

όπου το I είναι το διάστημα ανάλυσης πλαισίου. Η αυτοσυσχέτιση του πλαισίου ανάλυσης καθορίζεται από την σχέση

$$R[|k|] = \sum_{n=-\infty}^{\infty} s_i[n] * s_i[n + |k|] \qquad (2.10)$$

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

$$\mathbf{R*a = r} \quad (2.11)$$

όπου $a=\{a_1, a_2 \ldots a_P\}$ είναι το διάνυσμα των LPC συντελεστών, και **R** είναι ο πίνακας των συντελεστών αυτοσυσχέτισης και καθορίζεται ως

$$\mathbf{R}[i,j] = R[|i-j|] = \sum_{n=-\infty}^{\infty} s_i[n] * s_i[n + |k|] \qquad (2.12)$$

και $r=\{R[1],\ldots R[P]\}$. 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 του LPC συντελεστές. Αυτός ο αλγόριθμος μπορεί να συνοψισθεί από το παρακάτω σετ εξισώσεων:

$$\mathbf{E^0} = R[0] \qquad (2.13)$$

$$k_i = \left[R[i] - \sum_{j=1}^{i-1} a_j^{i-1} R[i-j]\right] / E^{i-j} \qquad (2.14)$$

$$a_i^i = k_i \qquad (2.15)$$

$$a_j^i = a_j^{i-1} + k_i * a_{i-j}^{i-1},\ 1 \le j \le i-1 \quad (2.16)$$

$$E^i = \left(1 - k_i^2\right) * E^{i-1} \quad (2.17)$$

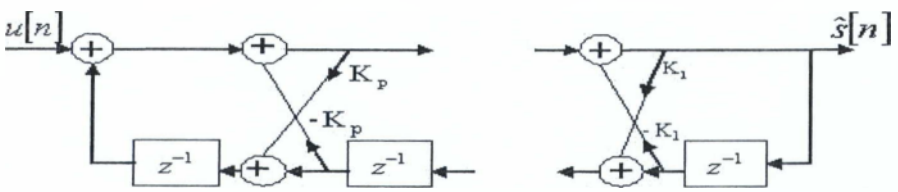
Οι εξισώσεις (2.13) και (2.14) λύνονται περιοδικά για $i=1,...,P$. Οι συντελεστές k_i για $i=1,...,P$ περιέχουν την ίδια πληροφορία με τους LPC συντελεστές, και ονομάζονται συντελεστές ανάκλασης (reflection coefficients) ή μερικοί συντελεστές συσχέτισης (γνωστοί και ως PARCORs). Το φίλτρο φωνητικού σωλήνα 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 (2.17) μας δείχνει ότι όλοι οι PARCOR συντελεστές έχουν μέγεθος λιγότερο από ένα. Έτσι

$$-1 \le k_i < 1, \quad (2.18)$$

Επειδή το LPC φίλτρο φωνητικού σωλήνα είναι περιοδικό, η σταθερότητα είναι ένα πρόβλημα. Αλλά όπως φαίνεται η συνθήκη της εξίσωσης (2.18) 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



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

ενέργεια του ελαχιστοποιείται. Έτσι η ποσότητα που καθορίζεται από

$$E = \sum_{-\infty}^{\infty} e^2[n] * w[n] \quad \textbf{(2.19)}$$

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

$$\Phi * \mathrm{a} = \varphi \quad \textbf{(2.20)}$$

$$\Phi[i,j] = \sum_{n=-\infty}^{i-1} s[n-i] * s[n-j] \quad \textbf{(2.21)}$$

και $\varphi = \{\Phi[1,0], \ldots, \Phi[P,0]\}$. Καθώς το Φ δεν είναι Toeplitz πίνακας δεν μπορεί να επιλυθεί τόσο αποτελεσματικά σε σχέση με τις κανονικοιημένες εξισώσεις της μεθόδο 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 του 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

$$1 - \frac{E_{p+1}}{E_p} < t_e \quad \textbf{(2.22)}$$

τότε μια καλή επιλογή είναι $P=p$.

Για ομιλία, δύο πόλοι (ένα πολικό ζεύγος) χρησιμοποιούνται για να μοντελοποιηθεί το κάθε formant.

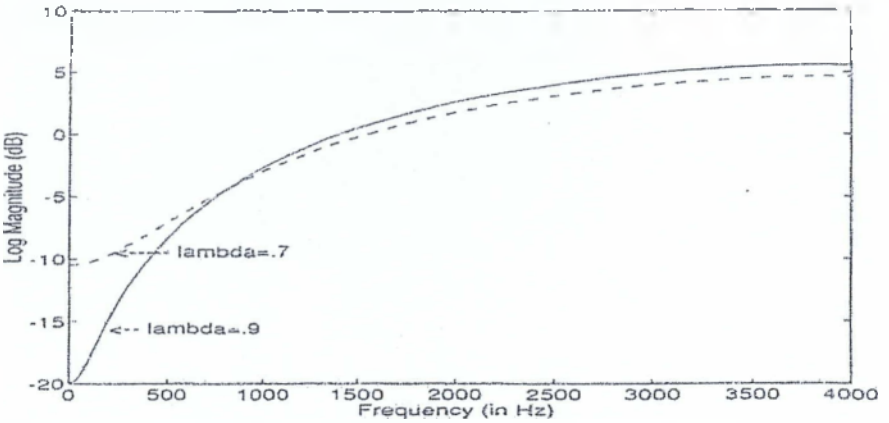
Το σήμα ομιλίας επίσης έχει φασματικά μηδενικά, αλλά επειδή αυτά έχουν ελάχιστες επιδράσεις, δεν μοντελοποιούνται στη συνάρτησης μεταφοράς του φωνητικού σωλήνα. Πρακτικά, για ομιλία 8 kHz, χρησιμοποιούνται τάξεις του προγνώστη σε ένα εύρος μεταξύ 10 και 16.

2.3.2 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

$$V_{pre}(z)=1-\lambda z^{-1}, \quad (2.23)$$

όπου $V(z)$είναι αποτελεσματικό ήπιο υψηπερατό 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.

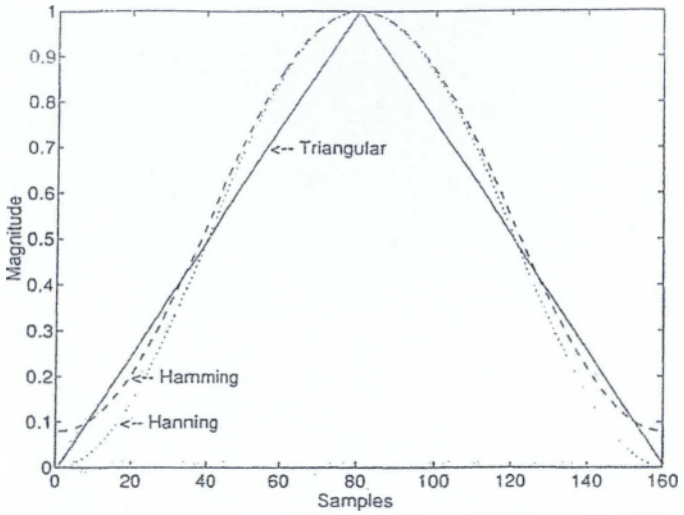
Για να εξουδετερώσουμε την επίδραση της προ-έμφασης, στο δέκτη έχουμε ένα αντίστοιχο φίλτρο από-έμφασης της μορφής

$$(z) = \frac{1}{1 - \eta z^{-1}}, \qquad (2.24)$$

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.

2.3.2 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: Οριμένα παράθυρα που χρησιμοποιούνται κατά την LPC ανάλυση. Το σχήμα δείχνει τα παράθυρα Hamming, Hanning και το τριγωνικό παράθυρο.

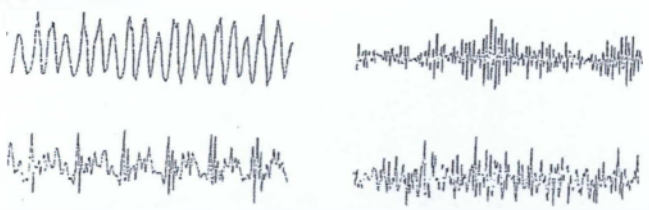
Το διάστημα πλαισίου ανάλυσης, *I*, καθορίζει τον αριθμό των δειγμάτων πάνω στα οποία θα χρησιμοποιηθούν οι LPC συντελεστές που προκύπτουν. Ο λόγος *I*/*L* αναπαριστά το ποσό της επικάλυψης μεταξύ δύο γειτονικών πλαισίων ανάλυσης (Σχήμα 2.5). Τυπικά χρησιμοποιείται μια διέγερση της τάξης του 50% (*I*=*L*/2). 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.

2.4 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. (δεξία) 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 (2.7) μπορεί να γραφτεί ως

$$s[n] = \sum_{i=1}^{p} a_i s[n-i] + e[n], \tag{2.25}$$

Συγκρίνοντας τις εξισώσεις (2.1) και (2.25) είναι φανερό ότι αν Gu[n]=e[n], το η έξοδος του H(z) θα είναι ίση με την αυθεντική ομιλία, δηλαδή s[n]=s[n]. 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 ψευδο-περιοδικό σήμα, ενώ για την άφωνη ομιλία είναι ένα σήμα που ομοιάζει με θόρυβο. Στους διεγερόμενους 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 αναλυώμενο πλαίσιο ομιλίας πρέπει να ταξινομηθεί ως έμφωνο ή άφωνο, και δεύτερον, για τα έμφωνα τμήματα πρέπει να καθοριστεί η περίοδος θεμελιώδους συχνότητας.

2.4.1 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.

2.4.2 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:

$$G = [R(0) - \sum_{k=1}^{p} a(k) * R(k)]^{1/2} \qquad (2.26)$$

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	l	40-160	120
de-emphasis factor	η	0.7-0.95	0.8

Πίνακας 2.2: Οι παράμετροι της LPC σύνθεσης.

2.5 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, … , P
2. Περίοδος θεμελιώδους συχνότητας
3. Κέρδος
4. Παράμετροι εμφώνων

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 πολυωνύμων $P(z)$ και $Q(z)$ που δίνονται από

$$P(z) = \left(1 - A(z)\right) + z^{-(P-1)} * (1 - A(z^{-1})) \qquad (2.27)$$

$$Q(z) = (1 - A(z)) - z^{-(P+1)}(1 - A(z^{-1})), \quad \textbf{(2.28)}$$

3. Τα πρώτα P δείγματα της κρουστικής απόκρισης των $H(z)$, $h(n)$.
4. Λόγοι λογαριθμικών περιοχών(LARs), που ορίζονται να είναι

$$\boldsymbol{LAR_i = log[(1 - k_i)/(1 + k_i)} \qquad \textbf{(2.29)}$$

5. Συντελεστές αυτοσυσχέτισης, $R[i]$'s
6. Συντελεστές Cepstrum του $h[n]$, οι οποίοι μπορούν να παρθούν από την περιοδικά επαναλαμβανόμενη

$$\boldsymbol{h(n) = a_n + \sum_{k=1}^{n-1}(\frac{k}{n} * \hat{h}[k]a_{n-k})} \qquad \textbf{(2.30)}$$

2.6 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 (2.11).

Εφαρμόζοντας τον z-μετασχηματισμό στην εξίσωση (2.7) έχουμε

$$E(z) = (1 - A(z))S(z), \quad \textbf{(2.31)}$$

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

$$\boldsymbol{E = \sum_n e^2[n] * \int_{-\pi}^{\pi} |E(j\omega)|^2} \quad \textbf{(2.32)}$$

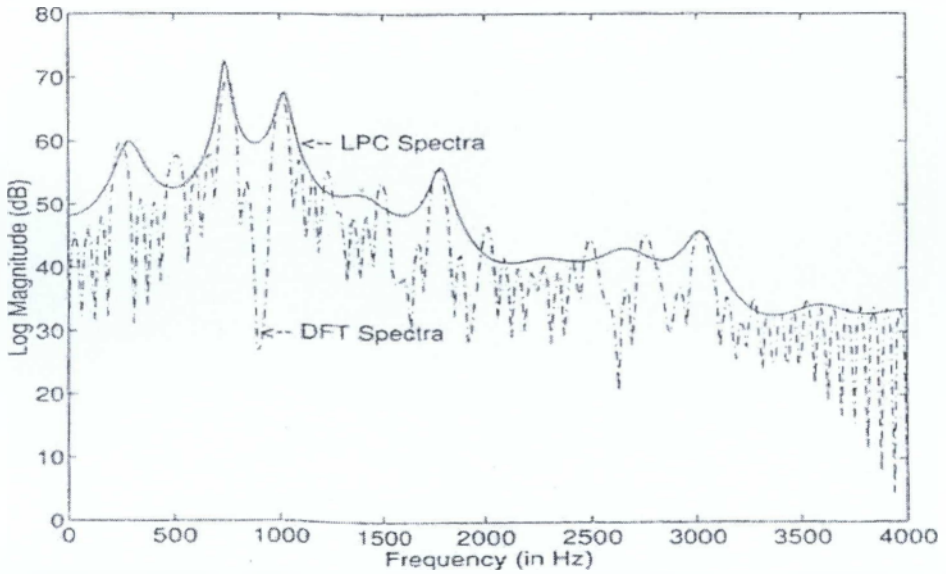
Συνδυάζοντας τις εξισώσεις (2.31) και (2.32), το **E** μπορεί να εκφραστεί ως

$$\boldsymbol{E = \frac{G^2}{2\pi}\int_{-\pi}^{\pi} \frac{|S(j\omega)|^2}{|H(j\omega)|^2}\, d\omega} \qquad \textbf{(2.33)}$$

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 (2.33) 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 $\left|H(e^{j\omega_i})\right|^2$ συνήθως ακολουθεί το φασματικό φάκελο του $\left|S(e^{j\omega_i})\right|^2$.

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.



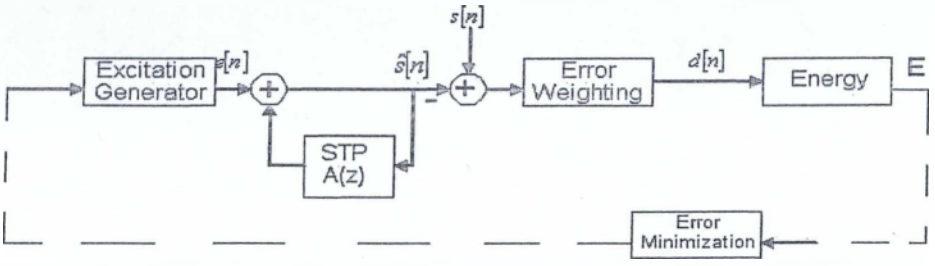
Σχήμα 2.10: Το FFT φάσμα και το 20-pole LPC φάσμα ενός τμήματος του σήματος ομιλίας.

ΚΕΦΑΛΑΙΟ 3

3.1 Εισαγωγή

Όλοι οι analysis-by-synthesis vocoders ανήκουν στην κατηγορία των διέγερσης κυματομορφής vocoders, και χρησιμοποιούν το μεγαλύτερο μέρος της διαθέσιμης πληροφορίας από το σήμα ομιλίας για την βελτίωση της ποιότητας και την ελαχιστοποίηση του ρυθμού μετάδοσης. Συγκεκριμένα, κάνουν χρήση της ακουστικής επικάλυψης θορύβου, της ακουστικής ανάλυσης συχνότητας, της ακουστικής αναισθησίας φάσης, μεταβολές ενέργειας συλλαβών, των ιδιοτήτων του φωνητικού σωλήνα μεγάλης-περιόδου, των ιδιοτήτων του φωνητικού σωλήνα μικρής-περιόδου, και πληροφορίες για τη θεμελιώδη συχνότητα κατά τη διάρκεια της διαδικασίας κωδικοποίησης. Γενικά, αυτό φέρνει τους analysis-bysynthesis κωδικοποιητές ανάμεσα στις τεχνικές κωδικοποίησης ομιλίας, με καλύτερη ποιότητα, χαμηλότερο ρυθμό μετάδοσης, και ιδιαίτερα απαιτητικές σε υπολογισμούς.

Οι analysis-by-synthesis κωδικοποιητές μπορούν να θεωρηθούν είτε ως διεγερώμενοι από τη θεμελιώδη συχνότητα LPCs, άλλα με ένα πιο αποτελεσματικό μοντέλο διέγερσης και μεγαλύτερο ρυθμό μετάδοσης, είτε ως διεγερώμενοι από την κυματομορφή LPCs που χρησιμοποιούν μια πολύ αποτελεσματική block-κωδικοποίησης τεχνική για την κωδικοποίηση του εναπομείναντος σήματος. Όλοι οι analysis- bysynthesis κωδικοποιητές αναπαριστούν το σήμα διέγερσης χρησιμοποιώντας ένα μικρό αριθμό από παραμέτρους συνήθως από 2 έως 6. Η διεργασία ανάλυσης διέγερσης, πραγματοποιείται συνθέτοντας ομιλία χρησιμοποιώντας κάθε δυνατό σετ παραμέτρων, και επιλέγοντας την καλύτερη συγκρίνοντας την με την αυθεντική ομιλία με τη χρήση μιας συνάρτησης που βασίζεται σε βάρη. Στους κωδικοποιητές analysis-bysynthesis, κάθε παράμετρος διέγερσης είναι ένας δείκτης σε ένα σύνολο συναρτήσεων διέγερσης. Έτσι λοιπόν, η διαδικασία analysis-bysynthesis είναι βασικά ένας τρόπος εξαντλητικής έρευνας σε ένα σύνολο από εναλλακτικές ακολουθίες για την εύρεση της βέλτιστης ακολουθίας διέγερσης.



Σχήμα 3.1: Block διάγραμμα της διαδικασία ανάλυσης που χρησιμοποιείται από τους analysis-by-synthesis γραμμικής πρόγνωσης κωδικοποιητές.

Ένα block διάγραμμα της διαδικασίας ανάλυσης για τους analysis-by-synthesis vocoders φαίνεται στο Σχήμα 3.1. Οι παράμετροι του μικρής-περιόδου προγνώστη παίρνονται μέσω της γραμμικής πρόγνωσης ανάλυση της ομιλίας όπως περιγράφηκε στο προηγούμενο κεφάλαιο. Στους analysis-by-synthesis κωδικοποιητές, η γεννήτρια διέγερσης είναι ικανή να παράγει K (συνήθως 64-1024) διαφορετικές ακολουθίες διέγερσης, $e_k(n)$. Η διαδικασία ανάλυσης παράγει όλα τα K δυνατά ξεχωριστά σήματα

ομιλίας, $\tilde{s}_k(n)$, αφαιρεί το αυθεντικό σήμα ομιλίας, και υπολογίζει την ενέργεια στο σήμα σφάλματος. Έτσι ο αναλυτής πρέπει να κάνει K ξεχωριστές διεργασίες σύνθεσης για να επιλέξει τη βέλτιστη ακολουθία διέγερσης. Με τους όρους που περιγράφει το μοντέλο στο Σχήμα 3.1 οι διάφοροι analysis-by-synthesis LPCs διαφοροποιούνται από το σύνολο των ακολουθιών που χρησιμοποιούνται για την αναπαράσταση της διέγερσης. Οι ακολουθίες διέγερσης μπορούν να γενικά να χωριστούν στους, διέγερσης παλμού όπως ο πολλαπλών παλμών διέγερσης LPC (MPLPC), διέγερσης θεμελιώδους συχνότητας όπως ο αυτοδιεγερώμενος vocoder (SEV), και διέγερσης κώδικα όπως ο διέγερσης κώδικα LPC (CELP).

3.2 Μοντέλο Διέγερσης

Η analysis-by-synthesis είναι μια διεργασία κωδικοποίησης όπου το σήμα διέγερσης καθορίζεται σε μια block-by-block βάση. Γενικά θεωρείται ότι το σήμα διέγερσης για κάθε block, μπορεί να είναι ένας συνδυασμός διαφορετικών στοιχείων διέγερσης όπως

$$e[n] = \sum_{k=1}^{M} \beta_k e_k[n], \qquad (3.1)$$

όπου $e_k[n]$ είναι το k^{th} στοιχείο διέγερσης. Για τους κωδικοποιητές που θα αναφερθούν παρακάτω, τα στοιχεία διέγερσης μπορεί να είναι ενός από τους παρακάτω τρεις τύπους: ένας παλμός, μια codebook ακολουθία, ή η έξοδος ενός μεγάλης-περιόδου (θεμελιώδους συχνότητας) προγνώστη. Ο αριθμός των ξεχωριστών στοιχείων διέγερσης, M, είναι συνήθως μικρός. Τυπικά παραδείγματα είναι ο CELP, ο οποίος έχει δύο στοιχεία (μια codebook ακολουθία και ένα μεγάλης-περιόδου προγνώστη), ο SEV, ο οποίος έχει ένα στοιχείο (ένα μεγάλης-περιόδου προγνώστη), και ο MELP, ο οποίος μπορεί να έχει από δύο έως οχτώ στοιχεία (όλα παλμοί). Κάθε στοιχείο διέγερσης καθορίζεται από ένα thδείκτη, γ_k, και ένα ανταποκρινόμενο κέρδος, β_k. Ο βέλτιστος δείκτης για το k στοιχείο, γ_k, καθορίζεται από την διαδικασία analysis-by-synthesis. Για ένα δεδομένο k και ένα δεδομένο σύνολο από K ακολουθίες διέγερσης, $F_k = \{f_\gamma[n], \gamma = 1, \ldots, K\}$, το $e_k[n]$ επιλέγεται να είναι το $f_\gamma[n]$ που ελαχιστοποιεί τη μέση τετραγώνου διαφορά μεταξύ των συνθετικών και αυθεντικών ακολουθιών ομιλίας. Ο βέλτιστος δείκτης, γ_k, και η τιμή του σχετιζόμενου κέρδους, β_k, εκπέμπονται ώστε έτσι η ακολουθία διέγερση να μπορεί να αναπαραχθεί από το δέκτη.

Οι παράμετροι που σχετίζονται με τα στοιχεία της ακολουθίας διέγερση παίρνονται με μια μέθοδο υποβέλτιστης ακολουθίας. Αυτό επιτυγχάνεται με την αφαίρεση των επιδράσεων των προηγούμενων στοιχείων διέγερσης από το αυθεντικό σήμα ομιλίας πριν πάρουμε τις παραμέτρους του επόμενου στοιχείου. Το κίνητρο γι' αυτή την υποβέλτιστη προσέγγιση είναι η απαγορευτική πολυπλοκότητα της εύρεσης των βέλτιστων παραμέτρων για όλα τα στοιχεία διέγερσης. Για παράδειγμα, το να βρούμε τις βέλτιστες παραμέτρους παλμού σε ένα πολλαπλών διεγέρσεων LPC, έχει ως αποτέλεσμα ενός σετ μη γραμμικών εξισώσεων στις θέσεις παλμού και πλάτους. Παρόλο που η μη γραμμικές εξισώσεις μπορούν να λυθούν με επαναληπτικές διεργασίες, αυτές συνήθως είναι πολύ πολύπλοκες και καθόλου πρακτικές για κωδικοποιητές πραγματικού χρόνου.

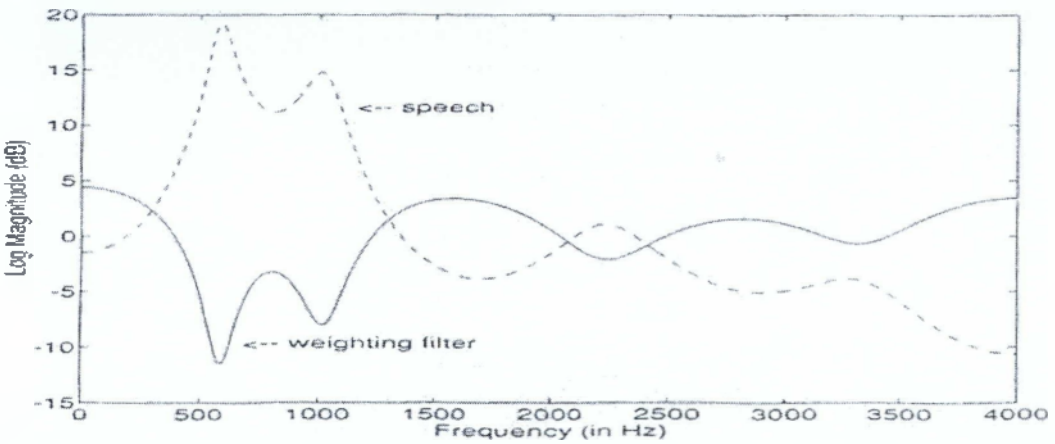
3.3 Error Weighting

Ένα σημαντικό στοιχείο του analysis-by-synthesis γραμμικής πρόγνωσης κωδικοποιητή είναι το error-weighting φίλτρο, που χρησιμοποιείται για να κατανείμει την ενέργεια της κωδικοποίησης του σήματος σφάλματος με τον κατάλληλο τρόπο. Η διαμόρφωση του φασματικού σφάλματος στα analysis-by-synthesis συστήματα προσπαθεί να καλύψει την κωδικοποίηση του σήματος σφάλματος (το καλυπτόμενο σήμα) με το σήμα ομιλίας (το σήμα που καλύπτει). Η αντιλαμβανόμενη ένταση της κωδικοποίησης του σήματος σφάλματος καθορίζεται από την ολική ισχύ του σήματος σφάλματος και τη φασματική κατανομή ως προς το αυθεντικό σήμα. Όταν το φάσμα θορύβου είναι επίπεδο, ο αντιληπτός θόρυβος είναι σε εκείνες τις περιοχές του φάσματος όπου η ομιλία έχει χαμηλή ενέργεια. Διαμορφώνοντας το φάσμα θορύβου έτσι ώστε να είναι ανάλογο του φάσματος σήματος μειώνει τον αντιληπτό θόρυβο, και έτσι βελτιώνει την γενική ποιότητα ομιλίας.

Σε αυτή την εφαρμογή του analysis-by-synthesis LPC, χρησιμοποιείται το error-weighting φίλτρο των Atal και Scroeber. Αυτό είναι ένα (βασισμένο στον LPC, block- adaptive weighting) φίλτρο που δίνεται από τη σχέση

$$W(z) = (1 - A(z))/[1 - A(a^{-1} * z)] \quad (3.2)$$

όπου το $0 \leq a \leq 1$ είναι η σταθερά βάρους (weighting), και ο όρος $A(z)$ είναι ο προγνώστης μικρής περιόδου. Το φίλτρο W(z) δίνει έμφαση στο σφάλμα στις κοιλάδες του φάσματος ομιλίας και μειώνει την έμφαση του στις περιοχές του formant. Το Σχήμα 3.2 δείχνει το φάσμα ενός έμφωνου τμήματος ομιλίας και την απόκριση έντασης του αντίστοιχου error-weighing φίλτρου |W(f)| για α=0.8.



Σχήμα 3.2: Το φάσμα ομιλίας για έμφωνο τμήμα και η απόκριση συχνότητας για το ανταποκρινόμενο error-weighting φίλτρο με α=0.8

Ένα σημαντικό πλεονέκτημα του error-weighting φίλτρου, $W(z)$, όπως καθορίζεται από την εξίσωση (3.2) είναι ότι τα μηδενικά του ακυρώνουν τους πόλους του LPC φίλτρου σύνθεσης. Η κρουστική απόκριση του συνδυασμένου φίλτρου είναι

$$H_\omega(z) = H(z)W(z) = \frac{1}{1 - A(\alpha^{-1}z)}, \qquad \textbf{(3.3)}$$

που αναφέρεται ως το LPC φίλτρο σύνθεσης με βάρη, μπορεί να προσομοιωθεί καλά από ένα σχετικά μικρό FIR φίλτρο,

$$h_\omega[n] \approx 0, \quad n > l, \qquad \textbf{(3.4)}$$

όπου το l είναι τυπικά 20. Το $h_\omega[n]$ αναφέρεται ως κρουστική απόκριση με βάρη (weighted impulse response). Μια τέτοια προσέγγιση έχει ως αποτέλεσμα τη μείωση των υπολογισμών στην analysis-by-synthesis διεργασία εύρεσης.

3.4 Διαδικασία Analysis-by-Synthesis

Υποθέτοντας ότι το μοντέλο analysis-by-synthesis πολλαπλών στοιχείων διέγερσης της εξίσωσης (1) χρησιμοποιείται το δείκτη και το κέρδος για κάθε στοιχείο διέγερσης. Το Σχήμα 3.3 δείχνει ένα block διάγραμμα αυτής της διαδικασίας ανάλυσης για μια γενική κατηγορία analysis-by-synthesis με πρόγνωση κωδικοποιητών. Κάθε στοιχείο της ακολουθίας διέγερσης παίρνεται με την ελαχιστοποίηση της ενέργειας του d[n], του οποίου ο z-μετασχηματισμός δίνεται από την σχέση

$$D(z) = Y(z) - \beta(\gamma)X_\gamma(z). \qquad \textbf{(3.5)}$$

Η $Y(z) = S(z)W(z)$ είναι ο z-μετασχηματισμός του αυθεντικού σήματος ομιλίας, και η $X_\gamma(z)$ είναι ο z-μετασχηματισμός της απόκρισης συτήματος (system response) στη δεδομένη ακολουθία διέγερσης, $f_\gamma[n]$. Το Σχήμα 3.3 δείχνει ένα block διάγραμμα για το πώς παίρνουμε το $d[n]$. Για κάθε δοσμένο γ, το ανταποκρινόμενο $\beta(\gamma)$, μπορεί να παρθεί με την ελαχιστοποίηση του σφάλματος μέσου τετραγώνου που δίνεται από τη σχέση

$$\boldsymbol{E_\gamma = \sum_{n=0}^{N-1} d^2[n] = \sum_{n=0}^{N-1}\{y[n] - \beta(\gamma) * \sum_{i=0}^{l-1} h_\omega[i] * f_\gamma[n-i]\}^2} \textbf{(3.6)}$$

όπου το $h_\omega[n]$ είναι η κρουστική απόκριση με βάρη (weighted impulse response). Εξισώνοντας την παράγωγο του E ως προς β του μηδενός, η εξίσωση για το $\beta(\gamma)$ δίνεται από τη σχέση

$$\boldsymbol{\beta(\gamma) = (\sum_{n=0}^{N-1} y[n] * x_\gamma[n])/(\sum_{n=0}^{N-1} x_\gamma{}^2[n])} \qquad \textbf{(3.7)}$$

όπου

$$x_\gamma[n] = \sum_{i=0}^{l-1} h_{\bar{\omega}}[i] * f_\gamma[n-i] \quad (3.8)$$

είναι η απόκριση συστήματος με βάρη (weighted system response) στη δοσμένη εξίσωση διέγερσης τα $e_k[n]$. Το σχετιζόμενο σφάλμα μέσου τετραγώνου δίνεται από

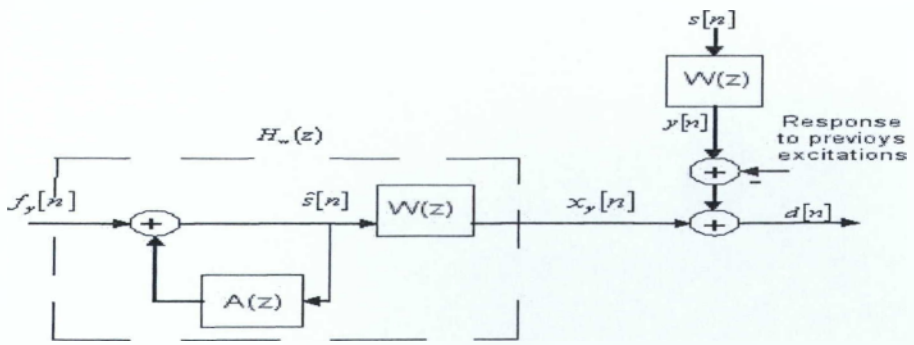
$$E_\gamma = (\sum_{n=0}^{N-1} y[n])^2 - \frac{(\sum_{n=0}^{N-1} x_\gamma[n] \cdot y[n])^2}{(\sum_{n=0}^{N-1} x_\gamma[n])^2} \quad (3.9)$$

Ο βέλτιστος δείκτης, γ_k, για τη συνάρτηση διέγερσης στοιχείου, $e_k[n]$, παίρνεται με την ελαχιστοποίηση του σφάλματος μέσου τετραγώνου, Ε, πάνω από τις επιτρεπόμενες τιμές του γ για τη συγκεκριμένη διέγερση παλμού που χρησιμοποιείται.

Δεδομένου οποιασδήποτε συνόλου διέγερσης, οι εξισώσεις (7)-(9), αποτελούν μια γενική μέθοδο για την επιλογή της ακολουθίας διέγερσης στοιχείων $e_k[n] = \beta_k f_{\gamma k}[n]$, που ελαχιστοποιεί την αντιλυπτή παραμόρφωση. Για να πάρουμε τις παραμέτρους των εναπομεινάντων στοιχείων διέγερσης, η απαραίτητη διεργασία είναι να απομακρύνουμε τα αποτελέσματα των προηγούμενων καθορισμένων στοιχείων της ακολουθίας, και να εκτελέσουμε την μέθοδο analysis-by-synthesis στο εναπομείναν σήμα.

3.5 Μεγάλης-Περιόδου Προγνώστες (Long-Term Predictors)

Στους γραμμικούς με πρόγνωση κωδικοποιητές, οι μικρής περιόδου (short-term) πλεονασμοί του σήματος ομιλίας (αυτοί εμφανίζονται λόγο του φαινομένου ακουστικού φιλτραρίσματος στο φίλτρο φωνητικού σωλήνα) απομακρύνονται με τη χρήση ενός μικρής-περιόδου προγνώστη. Όπως στους προσαρμοστικούς προγνωστικούς κωδικοποιητές. Όπως στους προσαρμοστικούς με πρόγνωση κωδικοποιητές, ο μεγάλης- περιόδου θεμελιώδους συχνότητας-περιοδικός πλεονασμός των εμφώνων τμημάτων του σήματος ομιλίας είναι εκμεταλλεύσιμος στους analysis-by-synthesis κωδικοποιητές. Το Σχήμα 3.4 μας δείχνει έναν LPC συνθέτη με δύο στη σειρά φίλτρα σύνθεσης. Ο μικρής-

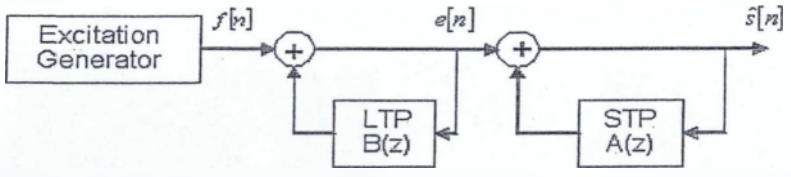


Σχήμα 3.3. Block διάγραμμα ενός μοντέλου πηγής ανάλυσης για την γενική κατηγορία των analysis-by-synthesis με πρόγνωση κωδικοποιητών, όπου s[n] είναι το σήμα ομιλίας στην είσοδο

περιόδου προγνώστης (STP), $A(z)$, μοντελοποιεί τη formant δομή στην κυματομορφή του σήματος, και ο μεγάλης-περιόδου προγνώστης (LTP), $B(z)$, μοντελοποιεί την αρμονική δομή της ομιλίας. Ορισμένες φορές, ο LTP αναφέρεται ως προγνώστης θεμελιώδους συχνότητας. Η γενική μορφή του προγνώστη μεγάλης-περιόδου είναι

$$B(z) = \sum_{i=-N_1}^{N_2} \beta_i z^{-\gamma-i}, \qquad \textbf{(3.10)}$$

όπου τα β_i είναι οι συντελεστές του μεγάλης-περιόδου προγνώστη (κέρδη), και γ είναι η καθυστέρηση του μεγάλης-περιόδου προγνώστη. Ο αριθμός των συντελεστών επιλέγεται να είναι από ένας έως τρεις. Η καθυστέρηση, γ, ενημερώνεται με τον ίδιο ρυθμό που ενημερώνονται και οι συντελεστές του προγνώστη.



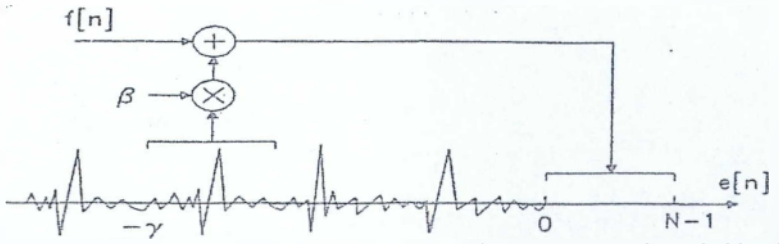
Σχήμα 3.4: Block διάγραμμα ενός γενικού analysis-my-synthesis LPC συνθέτη με μεγάλης-περιόδου προγνώστη.

Ο LTP μπορεί να θεωρηθεί ως μια πηγή διέγερσης του οποίου η έξοδος είναι ένα στοιχείο της συνολικής διέγερσης. Η γενική διέγερση μπορεί να εκφραστεί ως

$$e[n] = \beta e[n-\gamma] + f[n], \qquad \textbf{(3.11)}$$

όπου το $\beta e[n-\gamma]$ είναι το στοιχείο διέγερσης που παράγεται από τον LTP, και $f[n]$ είναι το άθροισμα των υπόλοιπων στοιχείων διέγερσης. Οι παράμετροι του μεγάλης- περιόδου προγνώστη παίρνονται από μια analysis-by-synthesis διεργασία στην οποία το γ είναι ο δείκτης και β είναι η ανταποκρινόμενη παράμετρος κέρδους. Το σύνολο αναζήτησης είναι ένα πεπερασμένο σετ από προηγούμενες ακολουθίες διέγερσης που μπορούν να αναπαρασταθούν από το σετ

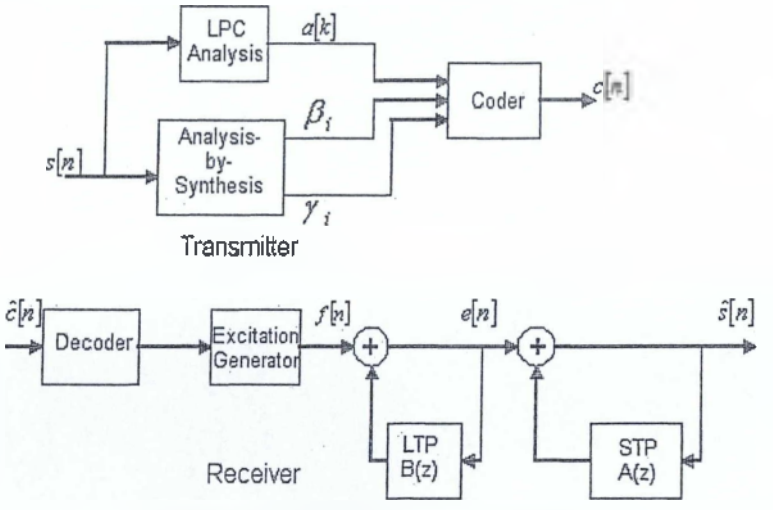
$$F = \{e[n-\gamma], \gamma = d_1, \mathrm{K}, d_2\}, \qquad \textbf{(3.12)}$$



Σχήμα 3.5: Search ensemble construction for the LTP. The optimum sequence is scaled by β and it is used to update the search ensemble.

όπου d_1 και d_2 καθορίζουν το εύρος των πιθανών τιμών καθυστέρησης που ανταποκρίνονται στο αναμενόμενο εύρος της θεμελιώδους συχνότητας στην ομιλία,

περίπου από 5 έως 20 msec. Το Σχήμα 3.5 μας δείχνει πως δημιουργείται το σύνολο αναζήτησης LTP μετά από κάθε πλαίσιο ανάλυσης. Αυτό το σύνολο στην πραγματικότητα είναι η μνήμη του LTP, και κάθε συνάρτηση συνόλου είναι μια ακολουθία N δειγμάτων που αρχίζουν από το δείγμα $n = -\gamma$. Έτσι η LTP καθυστέρηση, γ, είναι ο δείκτης ενός συνόλου του οποίου οι ακολουθίες δημιουργούνται ολισθαίνοντας ένα ορθογώνιο παράθυρο σε όλη τη μνήμη του LTP. Το κριτήριο για τη βέλτιστη καθυστέρηση είναι το κριτήριο του ελάχιστου σφάλματος με βάρη.



Σχήμα 3.6: Πομπός και δέκτης του MPLPC

3.6 LPC Πολλαπλών Παλμών Διέγερσης (MPLPC)

Στον πολλαπλών παλμών διέγερσης LPC (MPLPC), το σήμα διέγερσης μοντελοποιείται ως το άθροισμα των με βάρη μετατοπισμένων παλμών (weighted, delayed impulses). Το Σχήμα 3.6 δείχνει ένα απλό block διάγραμμα ενός MPLPC πομπού και δέκτη. Από την πλευρά του πομπού, το σήμα ομιλίας αναλύεται για να παράγει τους LPC συντελεστές, τις LTP παραμέτρους, και τις παραμέτρους πηγής. Οι παράμετροι πηγής αποτελούνται από τις θέσης του παλμού και των πλατών. Έτσι η πολλαπλών παλμών διέγερσης μπορεί να αναπαρασταθεί ως

$$f[n] = \sum_{k=1}^{M} \beta_k \delta[n - \gamma_k], \qquad (3.13)$$

όπου τα γ_k αναπαριστούν θέσεις παλμού και τα β_k είναι τα σχετιζόμενα κέρδη (ή πλάτη παλμού). Για τον MPLPC, η ολική ακολουθία διέγερσης δίνεται ως

$$e[n] = \beta_0 e[n - \gamma_0] + f[n] = \beta_0 e[n - \gamma_0] + \sum_{k=1}^{M} \beta_k \delta[n - \gamma_k], \qquad (3.14)$$

όπου τα β_0 και γ_0 είναι οι LTP παράμετροι και τα γ_k , β_k όπου $k = 1, \ldots, M$, είναι τα πλάτη και οι θέσης των M παλμών. Βασισμένοι στο μοντέλο διέγερσης της εξίσωσης (1), ο LTP και κάθε παλμός στο μοντέλο διέγερσης θεωρούνται να είναι ένα ξεχωριστό στοιχείο διέγερσης του οποίου οι παράμετροι (θέση και πλάτος) παίρνονται με τη χρήση της analysis-by-synthesis.

Στην analysis-by-synthesis της ομιλίας, καθορίζονται πρώτα οι LTP παράμετροι και η συνεισφορά του LTP αφαιρείται από το αυθεντικό σήμα ομιλίας. Έπειτα, αφού οι παράμετροι κάθε παλμού έχουν καθοριστεί, η συνεισφορά του αφαιρείται από το αυθεντικό σήμα και το υπόλοιπο σήμα χρησιμοποιείται για την εύρεση των παραμέτρων του επόμενου παλμού. Για σταθερού ρυθμού κωδικοποιητές, ο αριθμός των παλμών ανά πλαίσιο είναι συνήθως προκαθορισμένο. Είναι επίσης δυνατό να προσαρμόσουμε των αριθμό των παλμών ώστε να συναντήσουν ένα προκαθορισμένο κριτήριο σφάλματος, το οποίο έχει ως αποτέλεσμα ένα μεταβλητού ρυθμού κωδικοποιητή.

Οι παράμετροι του MPLPC αποτελούνται από τις παραμέτρους για την LPC ανάλυση (μήκος παραθύρου, *L*, μέγεθος πλαισίου LPC, *I*, τάξη προγνώστη, *P*), τον error weighting παράγοντα, *α*, το μέγεθος της πηγής ανάλυσης πλαισίου, *N*, και τον αριθμό των παλμών ανά πλαίσιο ανάλυσης, *M*. Ο Πίνακας 3.1 δίνει το εύρος και κάποιες τυπικές τιμές για αυτές τις παραμέτρους.

parameters	name	range	typical values
predictor order	*P*	1-16	10
LPC window length	*L*	160-360	240
LPC frame size	*I*	80-240	160
error-weighting factor	*a*	0-1	0.8
LTP (excitation) frame size	*N*	20-60	40
number of pulses exc. frame	*M*	2-20	8

Πίνακας 3.1: Οι παράμετροι του MPLPC ανάλυσης και σύνθεσης.

Μια υποκειμενική αποτίμηση των αποτελεσμάτων έχει δείξει ότι καλή ποιότητα ομιλίας μπορεί να παραχθεί από τον MPLPC σε ένα εύρος ρυθμού μετάδοσης από 9.6 έως 16 Kb/s. Η αποτελεσματικότητα ενός τέτοιου μοντέλου σε μέσους ρυθμούς μετάδοσης βρίσκεται στην ικανότητα του να μοντελοποιεί και την περιοδική δομή αλλά και την τυχαία φύση ενός LPC residual signal.

Οι παράμετροι που κωδικοποιούνται για μετάδοση αποτελούνται από τους LPC συντελεστές, την LTP καθυστέρηση και κέρδος, την θέση παλμού και τα πλάτη. Λόγο των παραμέτρων που κβαντίζονται, η απόδοση του MPLPC μειώνεται σημαντικά για ρυθμούς μετάδοσης κάτω από 9.6 kb/s. Για να μειώσουμε των αριθμό των παραμέτρων, οι θέσεις των παλμών μπορούν να σταθεροποιηθούν στο πλαίσιο ανάλυσης, το οποίο έχει ως αποτέλεσμα ενός μοντέλο τακτικού παλμού διέγερσης.

3.7 Τακτικού Παλμού-Διέγερσης LPC (RPLPC)

Ο RPLPC, είναι μια ειδική περίπτωση του μοντέλου πολλαπλών παλμών διέγερσης, στο οποίο οι παλμοί είναι ισοδύναμα κατανεμημένοι στο χρόνο. Στον RPLPC, ο πρώτος παλμός μπορεί να είναι βρίσκεται οπουδήποτε στο πλαίσιο ανάλυσης, αλλά ο αριθμός των παλμών σε ένα πλαίσιο και τα κενά μεταξύ τους είναι εξολοκλήρου

καθορισμένα. Έτσι, μετά τον καθορισμό του γ_1, οι θέσεις των υπολοίπων παλμών είναι $\gamma_1 \pm kD$ όπου το D είναι το κενό μεταξύ παλμών στην ακολουθία διέγερσης. Οι παράμετροι που κωδικοποιούνται για την τακτικού παλμού διέγερση είναι η θέση του πρώτου παλμού, γ_1, και τα πλάτη τωνπαλμών, β_k, $k = 1, \ldots, \frac{N}{D}$, όπου N, είναι το μήκος

της διέγερσης του πλαισίου ανάλυσης.

Η διαδικασία ανάλυσης περικλείει την ταυτόχρονη επιλογή της θέσης του παλμού και τις σχετιζόμενες τιμές κέρδους για την ελαχιστοποίηση του weighted error energy. Η θέση του πρώτου παλμού παίρνεται με τη χρήση της διαδικασίας analysis-by-synthesis, και όλα τα μεγέθη του παλμού παίρνονται από τη λύση ενός σετ από N/D γραμμικών εξισώσεων. Υψηλής ποιότητας ομιλία παίρνουμε με ένα διάστημα παλμών για $D = 4$ με ένα πλαίσιο ανάλυσης μεγέθους 40 στα περίπου 9.6 kb/s. Λόγο των προκαθορισμένων διαστημάτων μεταξύ των παλμών, ο RPLPC είναι υπολογιστικά λιγότερο πολύπλοκος από έναν συνηθισμένο MPLPC. Στον Πίνακα 3.2 τις παραμέτρους του RPLPC.

parameters	name	range	typical values
predictor order	P	1-16	10
LPC window length	L	160-360	240
LPC frame size	I	80-240	160
error-weighting factor	a	0-1	0.8
LTP (excitation) frame size	N	20-60	40
pulse seperation	D	2-10	4

Πίνακας 3.2: Οι παράμετροι του RPLPC ανάλυσης και σύνθεσης

3.8 LPC Διέγερσης Κώδικα (CELP)

Ο MPLPC αποδείχθηκε ότι είναι πολύ αποτελεσματικός κωδικοποιητής ομιλίας σε μεσαίους ρυθμούς μετάδοσης (για 9.6 kb/s και πάνω). Ένα μεγάλο τμήμα των bit στον MPLPC χρησιμοποιείται για να κωδικοποιηθούν οι παράμετροι διέγερσης, η θέση παλμού, και τα πλάτη. Ο αριθμός των bit για την κωδικοποίηση της διέγερσης πρέπει να μειωθεί σημαντικά για να επιτύχουμε χαμηλότερους ρυθμούς μετάδοσης. Για να μειωθεί ο ρυθμός μετάδοσης κάτω από τα 9.6 kb/s, πρέπει να μειωθεί περαιτέρω ο αριθμός των παραμέτρων διέγερσης. Το μοντέλο διέγερσης κώδικα (CELP) είναι πολύ αποτελεσματικό στη μοντελοποίηση της διέγερσης με πολύ μικρό αριθμό παραμέτρων. Ο διεγερόμενος από κώδικα γραμμικής πρόβλεψης κωδικοποιητή (CELP) είναι ένας ακόμα, βασισμένος στην analysis-by-synthesis, κωδικοποιητής ομιλίας για καλή ποιότητα σε χαμηλούς ρυθμούς μετάδοσης κάτω από 9.6 kb/s. Όπως και στον MPLPC, χρησιμοποιούνται προγνώστες μικρής και μεγάλης περιόδου για τη μοντελοποίηση του φασματικού φακέλου και της θεμελιώδους συχνότητας περιοδική δομή του σήματος ομιλίας.

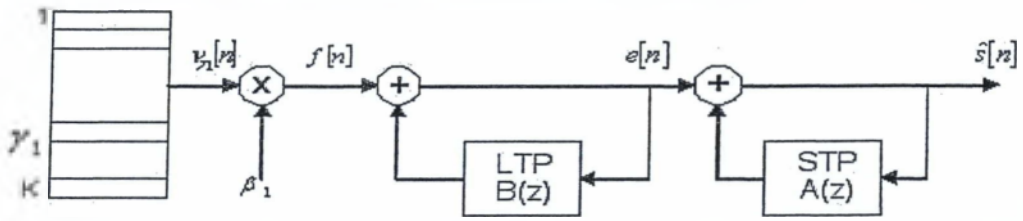
Στο μοντέλο διέγερσης κώδικα, ένα σχετικά μεγάλο codebook από τυχαίες ή ντετερμινιστικές ακολουθίες κώδικα (κωδικολέξεις) χρησιμοποιείται για τη μοντελοποίηση της διέγερσης. Στον CELP η συνολική διέγερση μπορεί να εκφραστεί ως

$$e[n] = \beta_0 e[n - \gamma_0] + \beta_1 v_\eta[n],$$

(3.15)

όπου $\upsilon_\gamma[n]$, $\gamma = 1,\ldots,K$, $n = 0,\ldots,N-1$, είναι το N-δείγμα ακολουθίας στο codebook μεδείκτη γ, όπου K είναι το μέγεθος του codebook. N είναι το μέγεθος του πλαισίου ανάλυση, το οποίο συνήθως επιλέγεται να είναι περίπου 5 msec. Αυξάνοντας το μέγεθος της ανάλυσης πλαισίου, έχουμε ως αποτέλεσμα τη μείωση της απόδωσης του κωδικοποιητή. Ένα τυπικό μέγεθος codebook είναι 1024, το οποίο απαιτεί $\log_2 1024 =$ 10bit για αναπαράσταση.

Στον πομπό, χρησιμοποιείται μια διαδικασία analysis-by-synthesis για να πάρουμε τη βέλτιστη κωδικολέξη. Η κωδικολέξη που προκύπτει από την ελαχιστοποίηση της ενέργειας του σφάλματος με βάρη (weighted error energy) επιλέγεται ως η βέλτιστη ακολουθία. Ο δείκτης της βέλτιστης κωδικολέξης και ο ανταποκρινόμενος παράγοντας κλίμακας κωδικοποιούνται και χρησιμοποιούνται για την παραγωγή της ακολουθίας διέγερσης στο συνθέτη. Το Σχήμα 3.7 είναι ένα block διάγραμμα ενός CELP συνθέτη.



Σχήμα 3.7. Block διάγραμμα ενός CELP συνθέτη.

Το κύριο μειονέκτημα του κώδικα διέγερσης είναι το υψηλό υπολογιστικό κόστος της διαδικασίας έρευνας. Το περισσότερο υπολογιστικό φορτίο στον CELP προέρχεται από εξαντλητική έρευνα του codebook, όπου τα φίλτρα σύνθεσης φιλτράρουν κάθε μια από τις υποψήφιες ακολουθίες. Για ένα codebook με 1024 εγγραφές και μήκος 40 απαιτούνται περίπου 500 εκατομμύρια διεργασίες πρόσθεσης-αφαίρεσης ανά δευτερόλεπτο. Πολλές διεργασίες έχουν προταθεί για αποτελεσματική έρευνα του codebook. Ορισμένες από αυτές, όπως η γρήγορης έρευνας με τη χρήση μοναδικής τιμής διάσπασης (SVD), έχει ως αποτέλεσμα τις βέλτιστες ακολουθίες με μια εξοικονόμηση στους υπολογισμούς. Πολλές άλλες διεργασίες παρέχουν υπολογιστική εξοικονόμηση εφαρμόζοντας μια μέθοδο προ-επεξεργασίας. Στην προ-επεξεργασία, ένα υποσέτ του codebook επιλέγεται και το επιλεγμένο υποσέτ ερευνάται διεξοδικά. Στην υλοποίηση του CELP μας, χρησιμοποιείται μια μεγάλη τυχαία ακολουθία αντί του codebook. Αυτή η μεγάλη ακολουθία μπορεί να ερευνηθεί αποτελεσματικά με τη χρήση ενός περιοδικά επαναλαμβανόμενου αλγόριθμου. Σε αυτό τον αλγόριθμο, η συνεισφορά του δείγματος στο τέλος της κωδικολέξης αφαιρείται, και η συνεισφορά του νέου δείγματος στην αρχή της κωδικολέξης προστίθεται στην απόκριση του φίλτρου. Ο Πίνακας 3.3 συνοψίζει τις παραμέτρους του CELP κωδικοποιητή και τις τυπικές τιμές τους.

parameters	name	range	typical values
predictor order	P	1-16	10
LPC window size	L	160-360	240
LP frame size	I	80-240	120
error-weighting factor	a	0-0.99	0.8
codeword (exc. Frame) size	N	20-60	40

Πίνακας 3.3: Οι παράμετροι ανάλυσης και σύνθεσης του CELP.

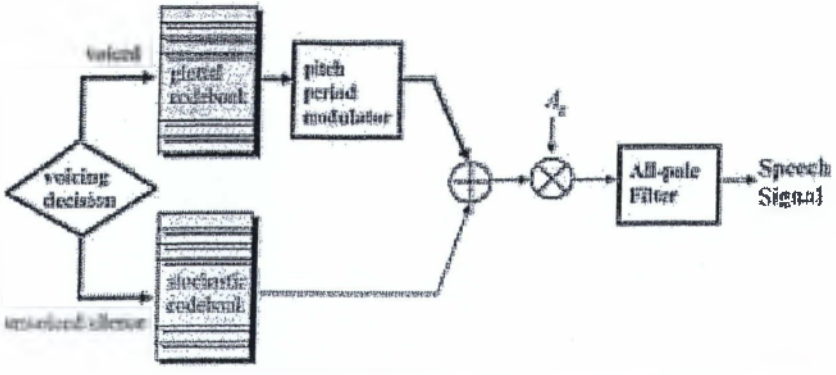
ΚΕΦΑΛΑΙΟ 4

4.1 Εισαγωγή

Εδώ γίνεται μια παρουσίαση ενός γλωττιδικά διεγερώμενου γραμμικής πρόβλεψης (GELP) κωδικοποιητή φωνής που κάνει κωδικοποίηση στα 1.9 kb/s. Ενώ η επεξεργασία των διαστημάτων της άφωνης ομιλίας και της σιωπής πραγματοποιείται με ένα στοχαστικό codebook, ένα γλωττιδικό codebook με 32 εισαγωγές για έμφωνη διέγερση χρησιμοποιείται για να προσομοιώσει τα χαρακτηριστικά που σχετίζονται με τη γλωττίδα. Περιγράφονται λεπτομερείς διεργασίες για την εξαγωγή των γλωττιδικών χαρακτηριστικών και για τη δημιουργία της διέγερσης. Εκτός από την μοντελοποίηση των χαρακτηριστικών φάσης, το αποτέλεσμα της γλωττιδικής διέγερσης παίρνει υπόψη του τη διασπορά παλμού και τον θόρυβο στροβίλισης. Ακροατές που πήραν μέρος στο τεστ μέσης γνώμης (mean opinion score, MOS) έδειξαν δυνατή προτίμηση προς τον GELP σε σχέση με τον 2.4 kb/s γραμμικής πρόβλεψης κωδικοποιητή (LPC-10e) παρόλο που η γενική απόδοση του GELP είναι κάπως κατώτερη σε σχέση με τον διέγερσης κώδικα γραμμικής πρόβλεψης (CELP) κωδικοποιητή.

4.2 Φάση Ανάλυσης

Για να μπορέσουνε να ενσωματώσουμε τη διέγερσης γλωττίδας στον LP κωδικοποιητή, υιοθετούμε ένα υβριδικό μοντέλο παραγωγής ομιλίας γραμμικής διέγερσης γλωττίδας (GELP). Όπως φαίνεται και στο Σχήμα 4.1, αυτό το μοντέλο έχει κληρονομήσει τη θεμελιώδη δομή του LPC κωδικοποιητή, στο οποίο η διέγερση μεταβάλετε μεταξύ εμφώνων και αφώνων σύμφωνα με τη φωνητική κατάσταση. Από την άλλη, η χρήση κωδικολέξης (codeword) αποκαλύπτει τη γενική φύση του CELP κωδικοποιητή. Τεχνικές που μπορούν να χρησιμοποιηθούν για την ανάπτυξη ολόκληρου του πλαισίου εργασίας εμπλέκουν την αναγνώριση των στιγμών κλεισίματος γλωττίδας (GCI), την εξαγωγή των χαρακτηριστικών φάσεως γλωττίδας, και τον καθορισμό του κέρδους της διέγερσης. Κάποιες τεχνικές που απαιτούνται μπορούν να δανειστούν από τους κωδικοποιητές LPC και CELP με τις απαραίτητες μετατροπές, και κάποιες πρέπει να αναπτυχθούν ανεξάρτητα για να συμβαδίσουν με τον κωδικοποιητή.

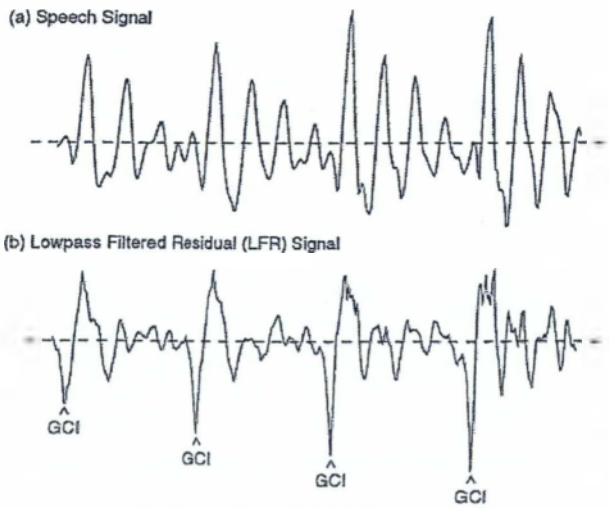


Σχήμα 4.1: Μοντέλο γλωττιδικής διέγερσης γραμμικής πρόγνωσης (GELP) παραγωγής ομιλίας.

Η επεξεργασία της άφωνης ομιλίας δεν απαιτεί περαιτέρω ανάλυση καθώς γι' αυτή χρησιμοποιείται ο τυπική CELP προσέγγιση. Από την άλλη η επεξεργασία των έμφωνων τμημάτων απαιτεί περισσότερη προσοχή. Όπως έχει αναφερθεί, πολλά γλωττιδικά χαρακτηριστικά βρίσκονται μέσα στα εναπομείναντα σήματα. Καθώς η ένταση του φάσματος του εναπομείναντος είναι θεωρητικά επίπεδη λόγο του αντίστροφού φιλτραρίσματος, η αναφορά του φάσματος φάσης στην προσωρινή πληροφορία είναι η βασική πηγή για την εύρεση των γλωττιδικών χαρακτηριστικών. Ενώ η γλωττιδική κίνηση παρουσιάζεται ως χαμηλής συχνότητας ροή αέρος κυματομορφή που βασίζεται στην περίοδο της θεμελιώδους συχνότητας, χρησιμοποιούμε μια σύγχρονης θεμελιώδους συχνότητας μέθοδο για την εξαγωγή των γλωττιδικών χαρακτηριστικών από το ολοκλήρωμα εναπομείναντος σήματος (residual). Το πρώτο πρόβλημα που παρουσιάζεται είναι ο καθορισμός των GSI's που οριοθετούν κάθε περίοδο θεμελιώδους συχνότητας. Έχει παρατηρηθεί ότι υπάρχει μεγάλη ομοιότητα μεταξύ του χαμηλοπερατά φιλτραρισμένου εναπομείναντος σήματος (LFR) και του γλωττιδικού διαφορικού παλμού. Έτσι λοιπόν είναι, πλεονεκτικότερο να εκτελέσουμε μια GCI αναγνώριση με τη χρήση του LFR. Το χαμηλοπερατό φίλτρο αποκτήσουμε το LFR έχει την παρακάτω μορφή:

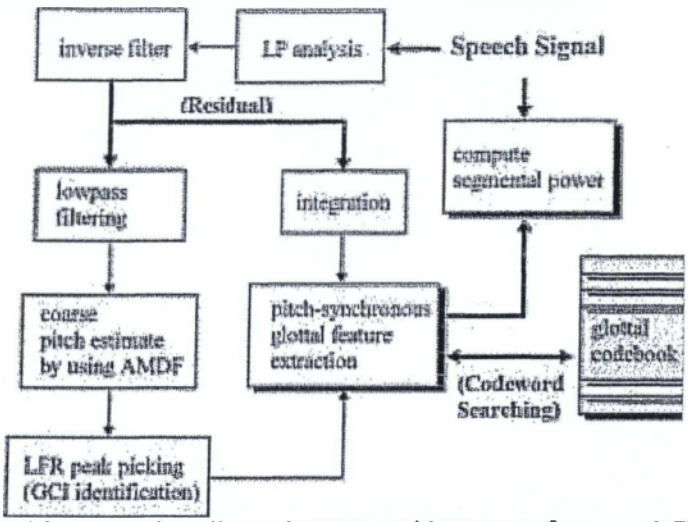
$$L(z) = \left(\frac{1}{1-0.95z^{-1}}\right)\left|\frac{1-z^{-1}}{\left(z-0.8z^{-1}\right)\left(1-0.5z^{-1}\right)}\right|^2 \quad \textbf{(4.1)}$$

Το σύνθετο φίλτρο μέσα στην πρώτη παρένθεση μιμείται τη φασματική μετατόπιση (spectral tilt) που χρησιμοποιείται για την εξαγωγή των διαφορικών γλωττιδικών παλμών δεδομένου ότι το υπόλοιπο τμήμα δηλώνει φίλτρο μηδενικής φάσης που χρησιμοποιείτε για τη μείωση του υψήσυχνου θορύβου και της ολίσθησης χαμηλής συχνότητας. Τροφοδοτώντας το LP εναπομείναν σήμα μέσω του $L(z)$, παίρνουμε ένα σήμα που μοιάζει με τον διαφορικό γλωττιδικό παλμό. Ένα τυπικό παράδειγμα του παραγόμενου LFR φαίνεται στο Σχήμα 4.2. Η GCI αναγνώριση είναι μια διαδικασία για την εύρεση μεγάλων αρνητικών κορυφών σε όλο το LFR.



Σχήμα 4.2: Παρουσίαση της GCI αναγνώρισης.

Ένα διάγραμμα ροής δίνεται το Σχήμα 4.3 που παρουσιάζει τις διεργασίες της εξαγωγής των γλωττιδικών χαρακτηριστικών. Η φάση της ανάλυσης ξεκινά με την απόφαση έμφωνου ή άφωνου καθώς και με τον προσδιορισμό της θεμελιώδους συχνότητας. Μέσα σε κάθε πλαίσιο, παίρνετε μια χονδρική τιμή της περιόδου της θεμελιώδους συχνότητας με την υιοθέτηση της συνάρτηση μέσης διαφοράς έντασης (AMDF) βασισμένη στο LFR. Το πλαίσιο καταχωρείται ως έμφωνω όποτε η μέση ένταση του σήματος ομιλίας είναι πάνω από 0.01 του μέγιστου ορίου



Σχήμα 4.3: Διεργασία ανάλυσης έμφωνης ομιλίας στον κωδικοποιητή GELP

και το αποτέλεσμα της AMDF παρουσιάζει μια διακριτή κοιλάδα γύρω από την περίοδο της θεμελιώδους συχνότητας. Σύμφωνα με αυτόν τον κανόνα διχοτόμησης, μια φωνή με ακανόνιστες μεταβολές θεμελιώδους συχνότητας θα καταχωρείται ως άφωνη ομιλία και η διέγερση που ανταποκρίνεται σε αυτή θα κωδικοποιείτε με μια μοναδική ακολουθία όπως γίνεται και στους κωδικοποιητές CELP. Πρέπει να σημειωθεί ότι ο διαχωρισμός μεταξύ εμφώνων/αφώνων δεν είναι απόλυτα κρίσιμος στον κωδικοποιητή GELP. Εκτός για φωνές με αφύσικες περιόδους θεμελιώδους συχνότητας, λάθος καταχωρημένα κομμάτια ομιλίας συχνά παρουσιάζουν περιοδικές αυθόρμητες διεγέρσεις σε σχέση με κάποια θορυβώδη στοιχεία. Καθώς η προσέγγιση CELP ανακτά μερικώς την περιοδικότητα μέσω της εύρεσης κωδικολέξεων, ούτε η γλωττιδική διέγερση ή ο στοχαστικός νεωτερισμός είναι αποδεκτός για τη σύνθεση αυτού του τύπου τεμαχίων ομιλίας.

Όταν το τεμάχιο ομιλίας κατηγοριοποιηθεί ως έμφωνο, προσαρμόζουμε την τιμή εκτίμησης της θεμελιώδους συχνότητας χρησιμοποιώντας ένα πλαίσιο «κοιτώντας προς τα μπροστά» (look-ahead) για να πάρουμε μια αξιόπιστη θεμελιώδη συχνότητα που ακολουθεί. Μια διεργασία επιλογής κορυφής ακολουθείται μετά, για την αναγνώριση του GCI κάτω από τον περιορισμό ότι το μέγεθος της απόκλισης μεταξύ δύο γειτονικών GCI's πρέπει να είναι λιγότερο από 20% της εκτιμιθέντας περιόδου θεμελιώδους συχνότητας. Μετά την ολοκλήρωση της αναγνώρισης του GCI, εφαρμόζουμε έναν συγχρονισμένο με τη θεμελιώδη συχνότητα αλγόριθμο για να εξάγουμε τα χαρακτηριστικά γλωττίδας που χαρακτηρίζονται από το ολοκλήρωμα του εναπομείναντος σήματος. Τα βασικότερα βήματα αποτελούνται, από την ολοκλήρωση,

την αφαίρεση της γραμμικότητας (linear trend removal), και την κανονικοποίηση μήκους. Εκτελούμε την ολοκλήρωση χρησιμοποιώντας ένα πρώτης τάξεως φίλτρο άπειρης κρουστικής απόκρισης (IIR), δηλαδή $1/(1-z^{-1})$, στο εναπομείναν σήμα που έχουμε πάρει. Η γραμμικότητα ως προς κάθε περίοδο θεμελιώδους συχνότητας, ορίζεται ως το διάστημα μεταξύ δύο συνεχόμενων GCI's και αφαιρείται για να διατηρηθεί κυκλική περιοδικότητα, κτλ.,

$$u(k) = u(k) + \frac{k}{N}\left(u(0) - u(N-1)\right), \quad k = 0,\ 1, \text{L},\quad N-1 \qquad \textbf{(4.2)}$$

όπου $u(k)$ αναπαριστά το ολοκληρωμένο residual με περίοδο N. Ο d.c. όρος απαλείφεται με το να αφαιρέσουμε το μέσο επίπεδο.

$$u(k) = u(k) - \bar{u}, \quad k = 0,\ 1,\ \text{L},\quad N-1 \qquad \textbf{(4.3)}$$

όπου το $\bar{u}$ δηλώνει το μέσο $u(k)$. Τέλος, το μήκος της περιόδου της θεμελιώδους συχνότητας παρεμβάλλεται γραμμικά για να πάρουμε ένα πρότυπο (template) 128 δειγμάτων. Το πρότυπο που προκύπτει από εδώ και πέρα θα ονομάζεται κανονικοποιημένο ολοκλήρωμα του εναπομείναντος σήματος (normalized integrated residual NIR).

Για να κωδικοποιήσουμε την ομιλία σεχαμηλού ρυθμούς μετάδοσης, απεικονίζουμε τη διέγερση πηγής σε μία λιτή έννοια χρησιμοποιώντας διανυσματικό κβαντισμό. Για το στάδιο της εκπαίδευσης του διανυσματικού κβαντισμού χρησιμοποιήθηκαν 6353 πρότυπα που πάρθηκαν από δώδεκα προτάσεις που εκφωνήθηκαν από τέσσερα άτομα. Κατά τη διεργασία εκπαίδευσης, το μέγεθος κάθε NIR προτύπου προσαρμόστηκε να παίρνει μονάδα. Η οποία είναι:

$$u(k) = u(k) \times \left(\frac{1}{128} \sum_{k=0}^{128} u^2(k) \right)^{-1/2}, \quad k = 0, 1, \text{L}, 127. \qquad \textbf{(4.4)}$$

Η μοναδιαία ισχύ είναι απαραίτητη για να διευκολυνθούν οι υπολογισμοί όταν υιοθετήσουμε την Ευκλείδεια απόσταση ως μέτρο της παραμόρφωσης στον διανυσματικό κβαντισμό. Το nth NIR πρότυπο $u_n(k)$ κατηγοριοποιείται στο σύμπλεγμα i

$$\sum_{k=0}^{128} u_n(k) v_i(k) > \sum_{k=0}^{128} u_n(k) v_j(k), \quad i \neq j; \quad j = 0, 1, \text{L}, J\text{-L} \qquad \textbf{(4.5)}$$

όπου J είναι ο αριθμός των συμπλεγμάτων, το $v_i(k)$ δηλώνει το κεντροειδές του ith συμπλέγματος (ή την γλωττιδική κωδικολέξη στην περίπτωση μας), που καθορίζεται το μέσο όλων των προτύπων στο ith υποσύνολο με την ενέργεια του να έχει προσαρμοσθεί ως μονάδα. Ο αλγόριθμος Linde-Buzo-Gray σε συνεργασία με το κριτήριο μέγιστης καθόδου (maximum descent) χρησιμοποιούνται για τη διάσπαση των συμπλεγμάτων. Ο κανόνας μέγιστης καθόδου διασπά τα συμπλέγματα, ένα κάθε φορά, για να μεγιστοποιηθεί η μείωση του αθροίσματος της παραμόρφωσης. Η διάσπαση

συμπλεγμάτων συνεχίζει έως ότου πάρουμε τον επιθυμητό αριθμό συμπλεγμάτων. Στην περίπτωση μας, 32 εισαγωγές βρέθηκε ότι μας δίνουν ικανοποιητική απόδοση για την περιγραφή του ολοκληρώματος του εναπομείναντος σήματος.

Εκτός από τα GCI's και τα γλωττιδικά χαρακτηριστικά φάσης, η ρύθμιση του κέρδους είναι το εναπομείναν θέμα στο στάδιο της ανάλυσης. Κωδικοποιούμε το λογάριθμο της ισχύος τμήματος αντί της διέγερσης κέρδους. Όπως θα διευκρινιστεί παρακάτω, η διέγερση κέρδους μπορεί να εξαχθεί από την ισχύ τμήματος χρησιμοποιώντας μια στρατηγική δύο φίλτρων.

4.3 Τρόπος Κωδικοποιήσης (Coding Scheme)

Δεδομένου ότι το σήμα ομιλίας δειγματοληπτίται στα 8 KHz, ενημερώνουμε το πλαίσιο ανάλυσης με ρυθμό 240 δειγμάτων. Ο σχεδιασμένος ρυθμός κωδικοποίησης είναι 1.9 Kb/s. Ο Πίνακας 4.1 παρουσιάζει ένα λεπτομερές τρόπο κωδικοποίησης. Για κάθε ξεχωριστό πλαίσιο, δύο σετ LP παραμέτρων are respectively εξάγονται από τα δείγματα ομιλίας που βρίσκονται στις θέσεις ένα τέταρτο και τρία τέταρτα του βασικού πλαισίου, τα καθένα από τα οποία εκτείνεται για μια διάρκεια 200 δειγμάτων. Και τα δύο μετατρέπονται σε παραμέτρους φασματικής γραμμής ζεύγη (LSP) και μαζί κωδικοποιούνται με τη χρήση ενός τεσσάρων-σταδίων 32-bit (π.χ., $\{8,8,8,8\}$) διανυσματικού κβαντισμού. Συγκεκριμένα, ομαδοποιούνται δύο σετ LSP παραμέτρων από επιτυχημένα πλαίσια για να παραχθεί ένα διάνυσμα. Αυτό το διάνυσμα κβαντίζεται από τον πρώτου-σταδίου διανυσματικό κβαντιστή, και το διάνυσμα λάθους που παράγεται κβαντίζεται από έναν δευτέρου-σταδίου κβαντιστή. Ομοίως οι τρίτου και τέταρτου-σταδίου κβαντιστές κβαντίζουν διαδοχικά τα διανυσματικά λάθη των προηγούμενων σταδίων. Τελικά η κβαντισμένη έκδοση του LSP διανύσματος παίρνεται από το άθροισμα των εξόδων των τεσσάρων-σταδίων.

Ανάλογα με τις φωνητικές συνθήκες, δύο μηχανισμοί έχουν σχεδιαστεί για να αντιμετωπιστεί επιτυχώς η πηγή διέγερσης. Η κωδικοποίηση της άφωνης ομιλίας είναι μια ευθεία μετατροπή του CELP. Το πλαίσιο προς ανάλυση χωρίζεται σε δύο υποπλαίσια, με μήκος 120 δείγματα το καθένα. Ο σχεδιασμός του codebook και η έρευνα των κωδικολέξεων ακολουθούν το Federal Standard (FS)-1016. Από την άλλη, οι κωδικοποιημένες παράμετροι για κάθε έμφωνο τμήμα αποτελούνται από τη λογαριθμική ισχύ τμήματος (segmental power), τη γλωττιδική κωδικολέξη, και τον αριθμό των GCI's που σχετίζονται με τη θέση του τελευταίου GCI. Ο σκοπός της κωδικοποίησης της τελευταίας θέσης του GCI παρά της περιόδου της θεμελιώδους συχνότητας γίνεται για να επιτύχουμε υψηλότερη ανάλυση συχνότητας για υψήσυχνα σήματα ομιλίας. Για να απεικονίσουμε τις μεταβολές της ενέργειας της έμφωνης ομιλίας, παίρνουμε την pitch- synchronous rms ισχύ τμήματος σε λογαριθμική κλίμακα και παρεμβάλλονται με τέσσερις αντιπροσωπευτικές τιμές για κάθε πλαίσιο. Αυτές οι τέσσερις τιμές κβαντίζονται διανυσματικά με τη χρήση 7 bit. Το άφωνο πλαίσιο κωδικοποιείτε με παρόμοιο τρόπο, αλλά μόνο οι στάθμες ισχύος που προέρχονται από δύο υποπλαίσια συμμετέχουν στον κβαντισμό, και μόνο 5 bits χρησιμοποιούνται.

	Analysis Method	Bits/Frame	Bit Rate
Spectrum	LPC: 10 order autocorrelation + Hanning window	two sets of LSP parameters jointly coded using four-stage, {8,8,8,8}, vector quantization	1066.7
Voiced/Unvoiced	Magnitude & AMDF	1	33.3
Excitation	voiced type: open loop	glottal codebook index: 5 GCI position: 7 GCI no.: 4 gain(+) : 7	766.7
	unvoiced type: closed loop	stochastic codebook index: 8 × 2 excitation polarity (+/-): 1 × 2 gain: 5	
Frame synchronization		1	33.3

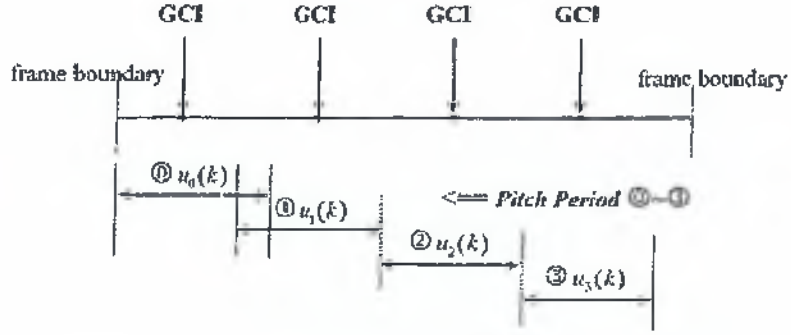
Note: Sampling Rate: 8 KHz; Frame Rate: 30 ms (240 Samples/Frame).

Πίνακας 4.1: Coding Scheme για τον 1.9 Kb/s GELP κωδικοποιητή.

Για κάθε πλαίσιο, χρησιμοποιούμε μόνο μια κωδικολέξη για να απεικονίσουμε τα χαρακτηριστικά πηγής καθώς η γλωττιδική κίνηση είναι σχετικά σταθερή παρά του ρυθμού με τον οποίο μεταβάλλετε η φωνητική περιοχή. Επιλέγουμε το ευρετήριο κωδικολέξεων c_n για να πάρουμε τη μέγιστη συσχέτιση σε ολόκληρο το πλαίσιο:

$$c_n = \arg\max_t \left\{ \sum_{k=0}^{127} \left(\sum_{m=0}^{N_{gci}} u_m(k) \right) v_t(k) \right\}, \quad (4.6)$$

όπου το N_{gci} αναπαριστά τον αριθμό των GCI's (ή αντίστοιχα, τις περιόδους της θεμελιώδους συχνότητας) στο παρών πλαίσιο. Καθορίζουμε τη θεμελιώδη συχνότητα ως το διάστημα που αρχίζει από μέσο μεταξύ του προηγούμενου και του τρέχοντος GCI έως το μέσο μεταξύ του τρέχοντος και του επόμενου GCI. Επειδή το αναλυόμενο πλαίσιο σπάνια περιέχει έναν ακριβή ακέραιο αριθμό περιόδων θεμελιώδους συχνότητας, και αλλάζουμε το αρχικό και/ή το τελευταίο διάστημα θεμελιώδους συχνότητας μέσα στο αναλυόμενο πλαίσιο ώστε να καλύψουμε ολόκληρη την περίοδο της θεμελιώδους συχνότητας. Το Σχήμα 4.4 παρουσιάζει το παράθυρο αλλαγής της περιόδου θεμελιώδους συχνότητας που μελετάμε. Τότε το GCI μέσα σε κάθε ανεξάρτητη περίοδο θεμελιώδους συχνότητας κυκλικά πηγαίνει στο αρχικό σημείο για να εκτελέσει μια έρευνα κωδικολέξεων.



Σχήμα4.4: Παρουσίαση κατακερματισμού μιας περιόδου θεμελιώδους συχνότητας σε ένα πλαίσιο.

4.4 Φάση Σύνθεσης

Ο αποκωδικοποιητής στο δέκτη είναι σχεδιασμένος να ανασυνθέτει σήματα ομιλίας από τα κωδικοποιημένα bit. Για τα άφωνα πλαίσια, η σύνθεση ομιλίας μπορεί εύκολα να επιτευχθεί διεγείροντας το φίλτρο σύνθεσης χρησιμοποιώντας στοχαστικές κωδικολέξεις προσαρμοσμένου κέρδους. Από την άλλη, η έμφωνη σύνθεση είναι μάλλον περίπλοκη γιατί πρέπει να αναπαράγουμε τα βασικά χαρακτηριστικά της έμφωνης διέγερσης βασιζόμενοι σε ημιτελής πληροφορίες. Καθώς η σύνθεση της άφωνης ομιλίας είδη έχει το χαρακτηριστικό του «ταιριάσματος» κυματομορφής (waveform matching), παρακολουθούμε τις μεταβολές της θεμελιώδους συχνότητας στα έμφωνα πλαίσια. Εκμεταλλευόμενοι τη συνδυασμένη πληροφορία που μας παρέχει η τελευταία θέση του GCI και ο αριθμός του GCI, μπορούμε να ανακτήσουμε τις εναπομείναντες GCI θέσεις στο βασικό έμφωνο πλαίσιο μέσω παρεμβολής και μαζί με επαρκή εξομάλυνση. Τότε πραγματοποιείται η σύνθεση ομιλίας κατά ένα συγχρονισμένο με τη θεμελιώδη συχνότητα τρόπο. Όπως έχει αναφερθεί, χρησιμοποιούμε μια μόνο κωδικολέξη για να χαρακτηρίσουμε τη γλωττιδική διέγερση για κάθε πλαίσιο. Παρόλο που σημαντικές διαφορές γλωττιδικής φάσης μπορούν να εμφανιστούν μεταξύ δύο γειτονικών περιόδων θεμελιώδους συχνότητας μέσα σε ένα πλαίσιο, αυτές οι ασυμφωνίες εμφανίζονται να έχουν μικρότερα αισθητό αποτέλεσμα στη συντεθειμένη ομιλία απ' ότι έχουν οι διαφορές στα όρια των πλαισίων. Έτσι, υπολογίζουμε κατά μέσο όρο την κωδικοποιημένη γλωττιδική διέγερση κατά ένα τρόπο που ομοιάζει με φίλτρο μέσα από

$$u_r(k) = 0.15^{1/N_{gci}} u_r(k) + \left(1 - 0.15^{1/N_{gci}}\right) u_c(k), \tag{4.7}$$

όπου $u_r(k)$ αναπαριστά το σύνολο των τρεχόντων διεγέρσεων και $u_c(k)$ είναι η επιλεγμένη κωδικολέξη για το τρέχον πλαίσιο. Η παραπάνω περιοδικά επαναλαμβανόμενη εξίσωση επιτρέπει στα γλωττιδικά χαρακτηριστικά προοδευτικά να ενσωματώνονται στη διέγερση. Κατά το στάδιο διεργασίας, επίσης, παρεμβάλουμε τις LSP παραμέτρους καθώς το συντιθέμενο διάστημα θεμελιώδους συχνότητας ολισθαίνει επί των πλαισίων.

Ας σημειωθεί ότι η αποκωδικοποιημένη γλωττιδική διέγερση είναι μια ακόμη ολοκληρωμένη έκδοση του εναπομείναντος σήματος. Παρά το γεγονός ότι η εφαρμογή ενός διαφοριστή υποβαθμίζει επαρκώς την επίδραση της ολοκλήρωσης, το συνολικό αποτέλεσμα της διέγερσης κυριαρχείται από στοιχεία χαμηλής συχνότητας λόγω της εξομάλυνσης από τη διανυσματική κβάντιση. Τέτοια συνέπεια, απαιτεί για μια διεργασία φασματικής προσαρμογής.

Αν υποθέσουμε ότι $v(n)$ είναι το διαφορικό αποτέλεσμα μιας γλωττιδικής κωδικολέξης με μήκος περιόδου *N*, πρώτα προσδιορίζουμε το

$$v(0) = \left(\frac{N+20}{120} \sum_{k=0}^{N-1} v^2(k) \right)^{1/2}. \tag{4.8}$$

Αυτή η τροποποίηση εισάγει μια θέση παλμού στο GCI. Η *n*th συνάρτηση Αυτοσυσχέτησης γίνεται

$$R(n)=\sum_{k=0}^{N-1} v(k)v((k+n).\mathrm{mod}.N),$$

(4.9)

όπου το σύμβολο **.mod** δηλώνει τις εργασίες των υποομάδων (modules). Κάθε αλλαγή σε ένα αυθαίρετο $v(i)$ θα έχει ως αποτέλεσμα διαφορετική συνάρτηση αυτοσυσχέτησης $\tilde{R}(n)$ όπως αυτή

$$\tilde{R}(n)=d(i)\left[v((i=n).\mathrm{mod}.N)+n((N+i-n).\mathrm{mod}.N)\right]+R(n), \quad (4.10)$$

όπου το $d(i)$ αναπαριστά την αλλαγή που υποτίθεται να αποφέρει ακολουθία αυτοσυσχέτισης ανταποκρινόμενο σε ένα αλλοιωμένο φάσμα. Ενώ η επιθυμητή διέγερση απαιτεί ένα επίπεδο φάσμα, σκόπιμα αλλοιώνουμε τα δείγματα γύρω από το GCI (π.χ., $i \ni \{2, 3, 4, N-1\}$) για να «ισοπεδώσουμε» το φάσμα της παραγόμενης διέγερσης.Θεωρητικά, οι συναρτήσεις αυτοσυσχέτησης για μια διέγερση με επίπεδο φάσμα θα πρέπει να είναι μηδέν εκτός από τον όρο μηδενικής υστέρησης (zero-lag). Παίρνοντας τις πρώτες p μη μηδενικές συναρτήσεις αυτοσυσχέτησης υστέρησης έχουμε

$$\begin{cases} d(i)\left[v((i+1).\mathrm{mod}.N)+v((N+i-1).\mathrm{mod}.N)\right]+R(1)=0, \\ d(i)\left[v((i+2).\mathrm{mod}.N)+v((N+i-2).\mathrm{mod}.N)\right]+R(2)=0, \\ \qquad \mathrm{M} \\ d(i)\left[v((i+p).\mathrm{mod}.N)+v((N+i-p).\mathrm{mod}.N)\right]+R(p)=0, \end{cases} \quad (4.11)$$

Αν $p>1$, το παραπάνω σετ εξισώσεων διαμορφώνει ένα υπερ-καθορισμένο σύστημα με τη λύση να είναι

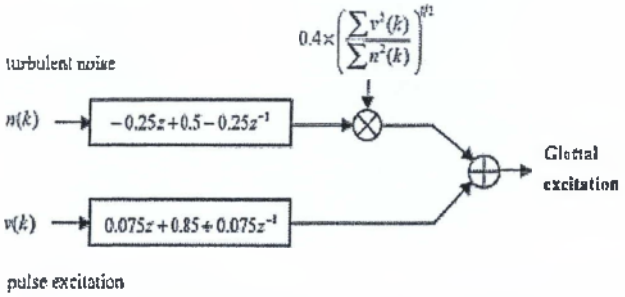
$$d(i)=-\frac{\sum_{n=1}^{p} R(n)\left[v((i+n).\mathrm{mod}.N)+v((N+i-n).\mathrm{mod}.N)\right]}{\sum_{n=1}^{p}\left[v((i+n).\mathrm{mod}.N)+v((N+i-n).\mathrm{mod}.N)\right]^2},$$

(4.12)

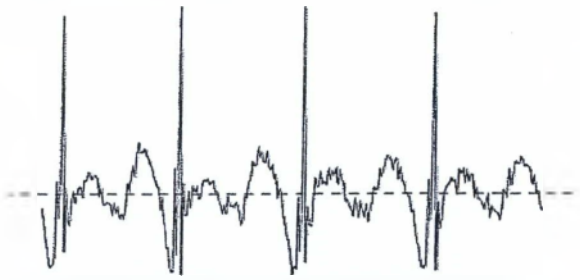
Καθώς ο γλωττιδικός θόρυβος στροβίλισης για την παραγωγή μιας φυσικής ακουστικά φωνής, το προτεινόμενο μοντέλο ενσωματώνει μια πηγή θορύβου με το φάσμα της διαμορφωμένο όπως φαίνεται στο Σχήμα 4.5. Παραδείγματα της παραγόμενης γλωττιδικής διέγερσης φαίνεται στο Σχήμα 4.6. Καθώς η παραγόμενη διέγερση έχει τα χαρακτηριστικά και της γλωττιδικής φάσης και του παλμού υψηλής ενέργειας, ονομάζουμε αυτή τη συνθετική διέγερση γλωττιδικό παλμό.

Ένα βοηθητικό όφελος που απορρέει από την παραπάνω διεργασία «ισοπέδωσης» φάσματος είναι η μείωση του βόμβου στη συνθετική ομιλία. Ο βόμβος στον παραδοσιακό LP vocoder γενικά αποδίδεται στην υψηλής συχνότητας

επιλεκτικότητα, που μπορεί να καθοριστεί από το λόγο του L1 κανόνα προς τον L2 κανόνα του σήματος διέγερσης.



Σχήμα 4.5: Εισαγωγή πηγής στροβιλοειδούς θορύβου.



Σχήμα 4.6: Παράδειγμα γλωττιδικής ώθησης.

Παρατηρούμε ότι ο γλωττιδικός παλμός προέρχεται από μια κωδικολέξη που προσομοιώνει τα χαρακτηριστικά χαμηλών συχνοτήτων του ολοκληρώματος του εναπομείναντος σήματος. Καθώς η φασματική«ισοπέδωση» της διέγερσης επιτυγχάνεται με την εισαγωγή ενός παλμού στο GCI και μετά προσαρμόζοντας τους γειτονικούς υποπαλμούς, είναι φανερό ότι το φάσμα υψηλής συχνότητας του παραγόμενου γλωττιδικού παλμού κυρίως αναδύεται από την ακολουθία παλμών. Σύμφωνα με τη διατύπωση του γλωττιδικού παλμού, ο παλμός που είναι προσδιορισμένος στο GCI κατέχει ένα μέγεθος που είναι ανάλογο προς τη γενική ενέργεια της διέγερσης κυματομορφής. Επιπλέον, η διεργασία φασματικής «ισοπέδωσης» διασπείρει αυτό τον παλμό υψηλής ενέργειας στον περιβάλλοντα χώρο. Ως αποτέλεσμα, η παραγόμενη επιλεκτικότητα είναι πολύ μικρότερη από αυτήν μιας καθαρής παλμικής διέγερσης. Η μείωση της επιλεκτικότητας στις υψηλές συχνότητες, με τη σειρά της, έχει ως αποτέλεσμα τη μείωση του βόμβου.

4.5 Προσαρμογή Κέρδους

Παρακολουθώντας την παραγωγή της γλωττιδικής πηγής και τους LSP συντελεστές, η συνθετική ομιλία εξασφαλίζεται τροφοδοτώντας την διέγερση προσαρμογής κέρδους σε ένα all-pole φίλτρο που απορρέει από τους LSP συντελεστές. Το κέρδος διέγερσης μπορεί να υπολογιστεί με διάφορους τρόπους. Λάθη στους υπολογισμού μπορούν να επηρεάσουν πολύ την ποιότητα της σύνθεσης. Για παράδειγμα, ενεργειακές μεταβολές στη συνθετική ομιλία μπορούν να θεωρηθούν ως κελάηδημα ή άλλα αντικείμενα. Στην κλασσική εργασία των Atal και Hanauer (1971), το κέρδος A_g προέκυπτε προσαρμόζοντας την αρχική ενέργεια, π.χ.,

$$P_r = \frac{1}{M}\sum_{k=0}^{M-1}\left(q(k)+A_g f(k)\right)^2 . \quad \textbf{(4.13)}$$

όπου $q(k)$ και $f(k)$ αναπαριστούν τη συμβολή της μνήμης των προηγούμενων πλαισίων και την απόκριση φίλτρου της παρούσας διέγερσης, αντίστοιχα. P_r είναι η τμηματική ενέργεια, M είναι το μέγεθος του πλαισίου. Το κέρδος εξάγεται από τη λύση της τετραγωνικής εξίσωσης. Αν το A_g είναι αρνητικό ή μιγαδικό, το θέτουμε ως μηδέν και για καθαρίσουμε τη μνήμη του φίλτρου. Όμως, μια μηδενική τιμή για τη διέγερση μπορεί να προκαλέσει διπλασιασμό της θεμελιώδους συχνότητας. Ο Hense Tohkura (1978) πρότεινε να μειωθεί η απόκριση φίλτρου ώστε η συμβολή της μνήμης να μπορεί να αγνοηθεί. Δηλαδή,

$$A_g = \left(MP_r \Big/ \sum_{k=0}^{M-1} f^2(k) \right)^{1/2} . \quad \textbf{(4.14)}$$

Ειρωνικά, αυτό μπορεί να έχει ως αποτέλεσμα μεταβολές στην ενέργεια εκτός και αν το φίλτρο είναι ιδιαίτερα υποβαθμισμένο, όμως ένα υπερβολικά υποβαθμισμένο φίλτρο τείνει να παράγει μια ένρινη ποιότητα.

Στην πραγματικότητα, τα μειονεκτήματα των παραπάνω δύο τεχνικών μπορούν να αποφευχθούν με τη χρήση της δύο-φίλτρων στρατηγικής. Παρατηρούμε ότι η μηδενική ρύθμιση χρησιμοποιείται μόνο ως μέσο εξασθένισης της ενέργειας του φίλτρου μνήμης. Μπορούμε να εξομοιώσουμε ένα τέτοιο αποτέλεσμα, μειώνοντας μόνο την απόκριση μνήμης και αφήνοντας τη διέγερση άθικτη. Συνεπώς, δύο φίλτρα παίρνουν μέρος στη διεργασία της σύνθεσης ομιλίας: αυτό που κρατάει τους συνηθισμένους LP συντελεστές είναι υπεύθυνο την απόκριση διέγερσης $f(k)$, και το μειωμένο φίλτρο τακτοποιεί τη συνεισφορά μνήμης $q(k)$. Επιτυγχάνουμε αυτή τη μείωση με το να πολλαπλασιάσουμε κάθε ένα από τους LP συντελεστές με το 0.97 υψωμένου εις τη δύναμη του δείκτη του. Κάτω από αυτές της συνθήκες, καμία από τις εξισώσεις (13) και (14) δεν υποφέρουν από τα παραπάνω προβλήματα. Επιπλέον, και οι δύο μπορούν να απορροφήσουν τις ενεργειακές αποκλείσεις λόγο του μετα-φιλτραρίσματος, το οποίο βοηθάει ώστε η συνθετική ομιλία να είναι εφάμιλλη της κυματομορφής της φυσικής ομιλίας σε περιοχές των formant. Το μετα-φιλτράρισμα δίνεται ως

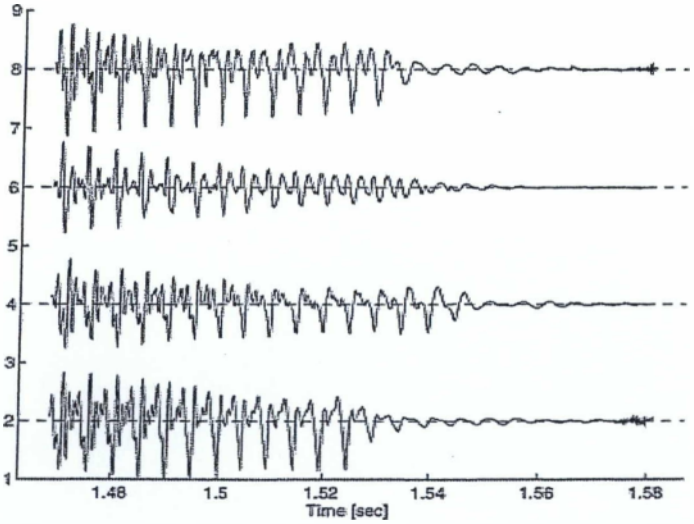
$$H(z) = \left(1 - 05z^{-1}\right) \frac{A(z/0.5)}{A(z/0.8)}, \quad \textbf{(4.15)}$$

όπου $A(z)$ είναι η συνάρτηση μεταφοράς του LP φίλτρου για το τρέχων πλαίσιο. Εδώ

προτιμούμε, να χρησιμοποιούμε την εξίσωσης (14) για να εξάγουμε το κέρδος καθώς απαιτεί λιγότερους υπολογισμούς. Όμως, επίσης βρίσκουμε ότι ένας πολλαπλασιαζόμενος παράγοντας μικρότερος του ένα είναι γενικά απαραίτητος για να εμποδίσουμε την αριθμητική αποκοπή. Αυτό έχει ως αποτέλεσμα, το κέρδος να υπολογίζεται ως

$$A_g = 0.95 \times \left(MP_r \Big/ \sum_{k=0}^{M-1} f^2(k) \right)^{1/2} . \qquad (4.16)$$

Στο Σχήμα 4.7 παρουσιάζουμε ένα τμήμα από ένα σήμα αυθεντικής ομιλίας μαζί με τις αποκωδικοποιημένες εκδοχές που παράγονται από τους LPC, CELP, GELP κωδικοποιητές, αντίστοιχα.



Σχήμα 4.7: Οι κυματομορφές από την πάνω προς τα κάτω είναι τα έμφωνα τμήματα μιας εκφωνήτριας και αποκωδικοποιημένη απόδοση με τη χρήση των LPC, CELP, και GELP κωδικοποιητών.

Όπως έχουμε δει ο κωδικοποιητής CELP, τείνει να ομοιάσει την κυματομορφή ομιλίας κατά μια χονδρική έννοια. Από την άλλη, ο LPC κωδικοποιητής διατηρεί τα φασματικά χαρακτηριστικά με παραμόρφωση της κυματομορφής ομιλίας. Ο GELP κωδικοποιητής παράγει μια κυματομορφή παρόμοια με το αυθεντικό σήμα κατά ένα συγχρονισμένο με τη θεμελιώδη συχνότητα τρόπο, εκτός από τις μεταβατικές περιοχές μεταξύ έμφωνων και άφωνων πλαισίων. Μια τέτοια ανακολουθία είναι αναπόφευκτη καθώς η σύνθεση ομιλίας πραγματοποιείται σε μία πλαίσιο προς πλαίσιο βάση. Για να παρέχουμε μια πιο ομαλή μετάβαση μεταξύ έμφωνης – άφωνης, υπάρχει μια επιπλέον περίοδος θεμελιώδους συχνότητας έμφωνης ομιλίας, που απλώνεται και μέσα στο άφωνο πλαίσιο. Δείγματα στο επιπλέον διάστημα θεμελιώδους συχνότητας παίρνονται από το άθροισμα με βάρη των εξόδων του φίλτρου από δύο διαφορετικούς τρόπους σύνθεσης. Οι συναρτήσεις βαρών για τους δύο αυτούς τύπους σημάτων συνθετικής ομιλίας επιλέγονται να είναι τραπεζοειδή παράθυρα με αντίθετες διευθύνσεις.

4.6 Συμπεράσματα

Εδώ περιγράφηκε ένας 1.9 Kb/s GELP κωδικοποιητής που κάνει χρήση του

μοντέλου γλωττιδικής διέγερσης για την έμφωνη ομιλία και καινοτόμων ακολουθιών για την άφωνη ομιλία. Αυτοί οι δύο τύποι διέγερσης σχηματοποιούνται σε codebooks που χρησιμοποιούνται για να διεγείρουν ένα κωδικοποιημένο all-pole φίλτρο που παίρνεται από παρεμβαλλόμενους LSP συντελεστές. Μόνο μία από τις δύο εξισώσεις διέγερσης είναι ενεργή κάθε στιγμή. Η απόδοση του έχει υποβληθεί σε υποκειμενικό τεστ σύγκρισης με τους LPC-10e (FS-1015) και CELP (FS-1016) κωδικοποιητές. Όπως βλέπουμε και στον Πίνακα 2 τα αποτελέσματα της μέσης γνώμης (MOS) δείχνουν ότι η απόδοση του παρόντος κωδικοποιητη είναι καλύτερη από τον LPC-10e vocoder και ελάχιστα χειρότερη από τον CELP κωδικοποιητή.

CODING SCHEME	CELP	GELP	LPC
Male speaker 1	3.43	3.70	1.91
Male speaker 2	3.55	3.57	2.02
Male speaker 3	3.20	3.28	2.19
Female speaker 1	3.21	2.59	1.92
Female speaker 2	3.53	3.25	2.46
Female speaker 3	3.17	2.45	1.91
AVERAGE	3.35	3.14	2.07

Πίνακας 4.2. Βαθμολογία Μέσης Γνώμης (MOS) για Συνθετικές Προτάσεις Ομιλίας με τη χρήση των CELP, GELP, LPC Κωδικοποιτών.

ΚΕΦΑΛΑΙΟ 5

Το συγκεκριμένο κεφάλαιο περιγράφει ένα νέο κωδικοποιητή Μικτής Διέγερσης Γραμμικής Πρόγνωσης (MELP), σχεδιασμένο για εφαρμογές πολύ χαμηλού ρυθμού μετάδοσης. Αυτός ο νέος κωδικοποιητής, μέσω αλγοριθμικών βελτιώσεων και βελτιωμένων τεχνικών κβάντισης, παράγει καλύτερη ποιότητα ομιλίας στα 1.7 kb/s σε σχέση με το νέο U.S Federal Standard MELP κωδικοποιητή στα 2.4 kb/s . Βασικά χαρακτηριστικά αυτού του κωδικοποιητή είναι ο βελτιωμένος αλγόριθμος εκτίμησης της θεμελιώδους συχνότητας και η LSF τεχνική κβαντισμού (Line Spectral Frequensies LSF) που απαιτεί μόνο 21 bits ανά πλαίσιο. Με κωδικοποίηση καναλιού, αυτός ο νέος MELP κωδικοποιητής είναι ικανός να διατηρεί καλή ποιότητα ομιλίας ακόμα και σε εξαιρετικά υποβαθμισμένα κανάλια, με ένα σύνολο ρυθμού μετάδοσης μόνο 3 kb/s.

5.1 Εισαγωγή

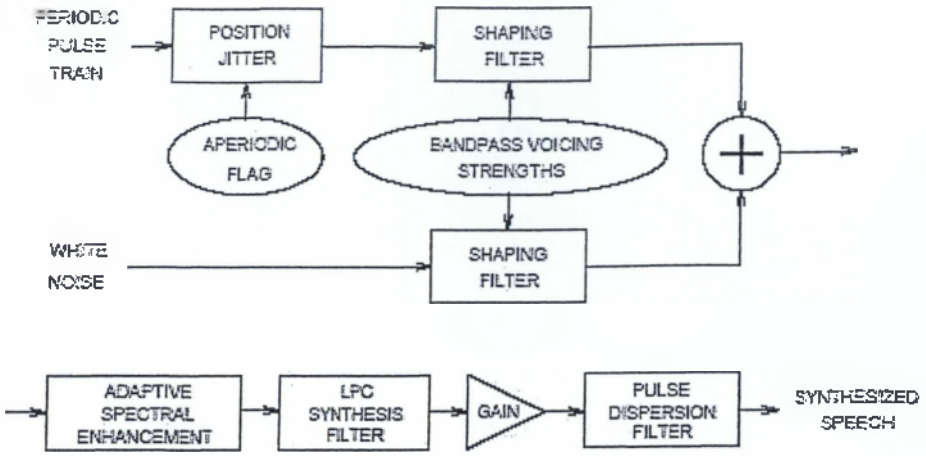
Ο κωδικοποιητής Μικτής Διέγερσης Γραμμικής Πρόγνωσης (MELP) υιοθετήθηκε ως το νέο U.S Federal Standard στα 2.4 kb/s. Παρόλο που τα 2.4 kb/s θεωρούνται γενικά χαμηλού ρυθμού μετάδοσης, υπάρχει ένας αριθμός εφαρμογών που ένας ακόμα χαμηλότερος ρυθμός μετάδοσης είναι απαραίτητος. Μία τέτοια εφαρμογή είναι η ασύρματη ψηφιακή μετάδοση ομιλίας, όπου κανάλια με φτωχό λόγο σήμα προς θόρυβο απαιτούν την εισαγωγή ενός υπολογίσιμου πλεονασμού bit ώστε να διατηρήσουν μια αποδεκτή ποιότητα ομιλίας, μειώνοντας έτσι τον αριθμό των bits που είναι διαθέσιμα στον κωδικοποιητή πηγής.

Εδώ περιγράφεται ένας κωδικοποιητής MELP που απαιτεί μόνο 1.7 kb/s και δίνει ανώτερη ποιότητα ομιλίας από τον κωδικοποιητή του Federal Standard στα 2.4 kb/s, τόσο για καθαρή όσο και για ενθόρυβη ομιλία. Κατάλληλα προστατευμένος με περιελικτηκούς κώδικες και με την ικανότητα να διαχειρίζεται σβησμένα πλαίσια, ο νέος αυτός MELP κωδικοποιητής είναι ικανός να διατηρεί τη βασική ποιότητα ακόμα και με την ύπαρξη ενός ποσοστού της τάξης του 5% τυχαίων λαθών.

5.2 Περιγραφή Κωδικοποιητή

Ο 1.7 kb/s MELP κωδικοποιητής βασίζεται, όπως και το νέο Federal Standard στο μοντέλο MELP. Αυτό το μοντέλο είναι βασισμένο στον παραδοσιακό LPC vocoder με είτε μια περιοδική εκπαίδευση παλμού, ή διέγερσης ενός all-pole φίλτρου από λευκό θόρυβο, αλλά περιέχει και τέσσερα επιπλέον χαρακτηριστικά. Όπως φαίνεται στο Σχήμα 5.1, ο συνθέτης έχει τις παρακάτω δυνατότητες: διέγερση με μίξη παλμού και θορύβου, περιοδικών ή απεριοδικών παλμών, προσαρμοστική φασματική βελτίωση, και ένα φίλτρο διασποράς παλμού.

Υπάρχουν τρεις σημαντικές διαφορές μεταξύ του 1.7 kb/s MELP κωδικοποιητή και του 2.4 kb/s Federal Standard: η βελτίωση του μοντέλου, η πιο αποδοτική κβάντιση και η κωδικοποίηση καναλιού.



Σχήμα 5.1: Συνθέτης MELP

5.3 Βελτιώσεις Μοντέλου

Βελτιώσεις στο μοντέλο MELP έχουν γίνει σε τρεις περιοχές. Πρώτον, έχουν βελτιωθεί ο καθορισμός της θεμελιώδους συχνότητας και των έμφωνων. Δεύτερον, ένα front-end καταστολής θορύβου έχει προστεθεί για να βελτιωθεί η απόδοση σε ακουστικά θορυβώδες περιβάλλον. Τέλος, το μέγεθος του πλαισίου έχει μειωθεί από 22.5 σε 20 ms. Αυτά έχουν ως αποτέλεσμα τη γενική αύξηση της ποιότητας ομιλίας.

5.3.1 Καθορισμός Θεμελιώδους Συχνότητας

Για τον καθορισμό της θεμελιώδους συχνότητας έχει αναπτυχθεί ένας αλγόριθμος βασισμένος στα υπο-πλαίσια που βελτιώνει σημαντικά την απόδοση σε σχέση με την προσέγγιση που χρησιμοποιείται στο Federal Standard και είναι βασισμένη στα πλαίσια. Ο αντικειμενικός σκοπός είναι να βρεθεί το ίχνος της θεμελιώδους συχνότητας σε ένα πλαίσιο ομιλίας, που ελαχιστοποιεί την ενέργεια που παραμένει από την πρόβλεψη της θεμελιώδους συχνότητας σε όλο το πλαίσιο, υποθέτοντας ότι χρησιμοποιείτε ο βέλτιστος συντελεστής πρόβλεψης της θεμελιώδους συχνότητας για κάθε μετατόπιση υπο-πλαισίου T_s. Αυτό το λάθος μπορεί να γραφεί ως άθροισμα N_s υπο-πλαισίων:

$$E = \sum_{s=1}^{N_s} E_s = \sum_{s=1}^{N_s} \left[\sum_n x_n^2 - \frac{\left(\sum_n x_n x_n - T_s\right)^2}{\sum_n x_n^2 - T_s} \right] \tag{5.1}$$

όπου x_n είναι το n δείγμα του σήματος εισόδου και το άθροισμα σε όλο το n περιλαμβάνει όλα τα δείγματα στο υπο-πλαισιο s. Η ελαχιστοποίηση αυτού του λάθους ισοδυναμεί με τη μεγιστοποίηση του κανονικοποιημένου συντελεστή συσχέτησης ρ που δίνεται από

$$\rho^2 = \frac{\sum_{s=1}^{N_s} \frac{\left(\sum_n x_n x_n - T_s\right)^2}{\sum_n x_n^2 - T_s}}{\sum_{s=1}^{N_s} \sum_n x_n^2} = \frac{\sum_{s=1}^{N_s} P_s \rho_s^2}{\sum_{s=1}^{N_s} P_s} \tag{5.2}$$

συσχέτησης μέσα στο υπο-πλαίσιο s. Στη συνέχεια επιβάλουμε ένα ίχνος θεμελιώδους συχνότητας εκμεταλλευόμενοι τον περιορισμό ότι κάθε εξασθένιση της θεμελιώδους συχνότητας σε κάθε υπο-πλαίσιο πρέπει να βρίσκεται μέσα σε μια συγκεκριμένη διακύμανση γύρω από μια γενική τιμή θεμελιώδους συχνότητας T:

$$\rho_s(T) = \max_{T_s = T-\Delta}^{T+\Delta} = \frac{\sum_n x_n x_n - T_s}{\sqrt{\sum_n x_n^2 \sum_n x_n^2 - T_s}} \tag{5.3}$$

όπου Δ είναι το ποσό της διακύμανσης της θεμελιώδους συχνότητας που επιτρέπεται ανάμεσα στα υπο-πλαίσια μέσα σε ένα πλαίσιο. Σημειωτέων ότι χωρίς τον εύρεσης της θεμελιώδους συχνότητας περιορισμό, το γενικό λάθος πρόγνωσης ελαχιστοποιείται με την εύρεση της βέλτιστης μετατόπισης, ανεξάρτητα για κάθε υπο-πλαίσιο. Επίσης, αυτή η μέθοδος διαφέρει από τη μέθοδο που βασίζεται στην αυτοσυσχέτηση στο ότι ενσωματώνει τις μεταβολές ενέργειας από το ένα υπο-πλαίσιο στο επόμενο.

Για τον καθορισμό της θεμελιώδους συχνότητας, μεταβάλουμε το T σε όλο το εύρος της θεμελιώδους συχνότητας βρίσκοντας την υψηλότερη κανονικοποιημένη συσχέτηση ρ του φιλτραρισμένου χαμηλοπερατά σήματος ομιλίας.

5.3.2 Καταστολή Θορύβου

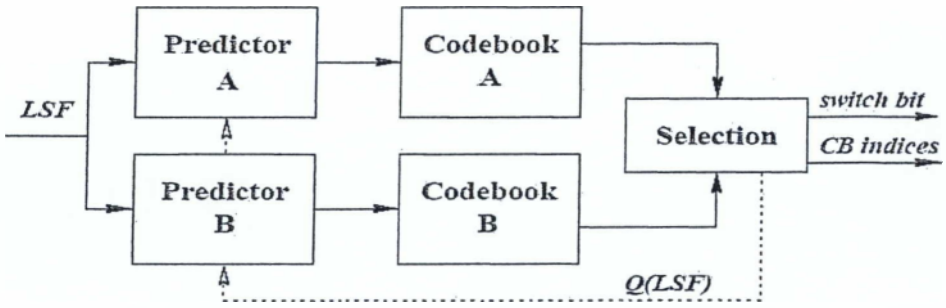
Η μέθοδος καταστολής θορύβου Smouthed Spectral Subtraction (SSS) που χρησιμοποιείται, βασίζεται στην παραδοσιακή φασματική αφαίρεση, όπου μια εκτίμηση του φάσματος της ισχύς θορύβου αφαιρείται από το φάσμα της θορυβώδους ομιλίας, αλλά έχει τρεις ξεχωριστές βελτιώσεις. Πρώτον, ένας περιοριστής χρησιμοποιείται στο φίλτρο καταστολής θορύβου $H(\omega)$ ώστε να μη μπορεί να πάει κάτω από μια ελάχιστη τιμή της τάξης των -10dB. Αυτό εμποδίζει το φίλτρο καταστολής θορύβου να έχει μια διακύμανση γύρω από πολύ μικρές τιμές κέρδους, και επιπλέον μειώνει την ενδεχόμενη παραμόρφωση του σήματος ομιλίας. Δεύτερον, η εκτίμηση της ισχύος του φάσματος του θορύβου αυξάνεται τεχνητά κατά ένα μικρό όριο (5dB) ώστε μικρά λάθη σε εκτιμήσεις στο φάσμα θορυβώδους σήματος να μην οδηγούν σε διακειμενόμενες εξασθενήσεις. Τρίτον, αντί να γίνεται χρήση των εκτιμήσεων που προέρχονται από τον FFT της θορυβώδους ομιλίας και του φάσματος του θορύβου κατευθείαν στον κανόνα εξασθένησης, γίνεται χρήση μιας εκδοχής της φασματικής ισχύος που έχει εξομαλυνθεί. Γίνεται χρήση μετακινούμενης μέσης εξομάλυνσης στη συχνότητα, ένα παράθυρο εξομάλυνσης μεγέθους 32 (για έναν FFT μεγέθους 256) βρέθηκε να δουλεύει καλά. Αυτή η εξομάλυνση ελαττώνει τις διακυμάνσεις των φασματικών εκτιμήσεων, η οποία εμποδίζει την εμφάνιση μουσικών θορύβων. Ως συνδυαζόμενο αποτέλεσμα αυτών των τριών βελτιώσεων, ο αλγόριθμος SSS είναι ικανός να εξασθενήσει τον ακουστικό θόρυβο περιβάλλοντος κατά 10 dB χωρίς να εισάγει καθόλου μουσικό θόρυβο.

5.4 Κβαντισμός

Η μεγάλη μείωση στο ρυθμό μετάδοσης σε αυτόν το νέο MELP κωδικοποιητή προέρχεται από τη νέα μέθοδο κβαντισμού LSF, που μειώνει τον αριθμό των bits που χρειάζονται για να αναπαρασταθεί το LPC φίλτρο από 25 σε 21 bits, χωρίς να υπάρχει κόστος στην αποθήκευση ή στην πολυπλοκότητα. Αποτελεσματικότερη κβάντιση της θεμελιώδους συχνότητας, των έμφωνων, και του κέρδους, μας εξασφαλίζουν τρία επιπλέον bit ανά πλαίσιο. Για να μειώσουμε το γενικό ρυθμό μετάδοσης, τα μεγέθη των σειρών Fourier που μεταδίδονται στον κωδικοποιητή του Federal Standard εδώ εξαλείφονται, «γλιτώνοντας» οχτώ bit ανά πλαίσιο.

5.4.1 LSF Κβάντιση

Εδώ έχει σχεδιαστεί ένας 21-bit με επιλογή πρόγνωσης κβαντιστής (switched predictive quantization) με καλύτερη απόδοση από των κβαντιστή των 25 bit που χρησιμοποιείται από το Federal Standard. Η περισσότερη από αυτή τη βελτίωση της αποτελεσματικότητας είναι λόγο της χρήσης προγνωστικού κβαντισμού, αλλά επιπλέον κέρδος στην απόδοση προκύπτει από τη χρήση της θεωρητικά καταλληλότερης LSF συνάρτησης βαρών.



Σχήμα 5.2: Block Διάγραμμα Switched-predictive LSF Κβαντιστή.

Χρησιμοποιούμε ένα με επιλογή-πρόγνωσης (switch predictive) πολλών σταδίων διανυσματικό κβαντισμό (MSVQ) από τους LSF's, όπως φαίνεται στο Σχήμα 5.2 . Για κάθε πλαίσιο ομιλίας, και τα δύο ζεύγη προγνώστη/codebook δοκιμάζονται, και αυτό που παρέχει την καλύτερη απόδοση κβαντισμού επιλέγεται για μετάδοση μαζί με ένα bit που αναπαριστά την πληροφορία για την επιλογή. Υπάρχει ένα σημαντικό πλεονέκτημα στη χρήση δύο διαφορετικών codebooks σε σχέση με τη χρησιμοποίηση ενός διαμοιραζόμενου codebook, χωρίς καμιά αύξηση στη πολυπλοκότητα σε σχέση με την περίπτωση που δεν εμπεριέχει πρόγνωση.. Η χρησιμοποίηση ξεχωριστών codebooks επιτρέπει στο καθένα να βελτιστοποιηθεί ξεχωριστά. Καθώς και τα δύο 4-σταδίων, 20 bit MSQV codebooks έχουν λιγότερο από το μισό μέγεθος, σε σχέση με την 25 bit με μη ύπαρξη πρόγνωσης έκδοση, Τόσο η αποθήκευση όσο και η πολυπλοκότητα αναζήτησης στην πραγματικότητα μειώνονται σε αυτό το νέα μέθοδο, και μπορούμε να αυξήσουμε το βάθος αναζήτησης της M-καλύτερης MSQV αναζήτησης από $M = 8$ σε $M = 12$ για ισοδύναμη πολυπλοκότητα.

Για εκπαίδευση, χρησιμοποιούμε μια επέκταση της επαναληπτικής MSQV

διαδικασία εκπαίδευσης, στην οποία εναλλασσόμαστε μεταξύ της εκπαίδευσης των συντελεστών πρόγνωσης, δεδομένου του codebook, και της εκπαίδευσης του codebook, δεδομένων των συντελεστών πρόγνωσης. Αυτός ο κλειστού-κύκλου μηχανισμός εναλλαγής εμπεριέχεται επίσης στη διαδικασία εκπαίδευσης Αυτό υλοποιεί μια πλήρη κλειστού κύκλου βελτιστοποίηση, και για τους συντελεστές του προγνώστη και για το codebook.

Επιπλέον της επιλογικής-πρόβλεψης, γίνεται χρήση μιας νέας LSF συνάρτησης βαρών για να γίνει προσέγγιση της φασματικής διαταραχής της συχνότητας με τη χρήση βαρών ($SD_{f\omega}$) που καθορίζεται από τη σχέση

$$SD_{f\omega}\left(A_q(z), A(z)\right) = \sqrt{\frac{1}{W_0}\int_{f=0}^{4000} |W_B(f)|^2 10\log_{10}\frac{|A_q(z)|^2}{|A(f)|^2} df} \tag{5.4}$$

όπου $A_q(z)$ και $A_q(z)$ παριστάνουν τα κβαντισμένα και μη κβαντισμένα LPC φίλτρα, W_0 είναι μια σταθερά κανονικοποίησης, και η Bark συνάρτηση βαρών $W_B(f)$ είναι καθορισμένη από

$$W_B(f) = \frac{1}{25 + 75(1 + 1.4\left(\frac{f}{1000}\right)^2)^{0.69}} \tag{5.5}$$

Προηγουμένως είδαμε ότι αυτή η συνάρτηση βαρών που βασίζεται στην κλίμακα Bark προβλέπει καλύτερα την προτίμηση του ακροατή στον κωδικοποιητή MELP, και τώρα παρουσιάζεται μια LSF συνάρτηση με βάρη η οποία βελτιστοποιεί αυτή τη μορφή του SD.

Σε υψηλούς ρυθμούς, τα βέλτιστα LSF βάρη για την ελαχιστοποίηση του SD χωρίς βάρη είναι ο πίνακας ευαισθησίας των LSFs:

$$\left.\frac{\partial^2 SD(a(\omega), a(\bar{\omega}))}{\partial\omega_k \partial\omega_i}\right|_{\omega=\bar{\omega}} = 4\beta j_{\omega_k}^T R_A j_{\omega_i} \tag{5.6}$$

όπου j_{ω_i} είναι η kth στήλη του Jacobian πίνακα για τα LSFs, R_A είναι ο πίνακας αυτοσυσχέτησης της απόκρισης παλμού του LPC φίλτρου σύνθεσης, και β είναι ένας παράγοντας κλίμακας. Στην πράξη, χρησιμοποιούμε μία προσέγγιση $8^{ης}$ τάξης ενός all-pole μοντέλου της συνάρτηση Bark με βάρη $W_B(f)$.

Κβαντιστής	$SD_{f\omega}$ (db)	>2dB (persent
25-bit	1.06	2.4
21-bit switched	0.97	0.81

Πίνακας 5.1: Απόδοση LSF κβαντιστή για χωρίς εξάρσεις ομιλία.

Στον Πίνακα 5.1 παρουσιάζεται η φασματική παραμόρφωση με βάρη για τον Federal Standard κβαντισμό και την με επιλογή πρόγνωσης εκδοχή. Αυτό το τεστ σετ μία χωρίς εξάρσεις ομιλία που δεν συμπεριλαμβανόταν στα σετ εκπαίδευσης. Ο 21-bit με επιλογή πρόγνωσης (switched-predictive) κβαντιστής είναι καθαρά ανώτερος από την με 25-bit χωρίς πρόγνωση εκδοχή, και σε όρους μέσης διαταραχής. Επίσης παρατηρούμε ότι και ακόμα για ιδιαιτέρως φιλτραρισμένη ομιλία, η οποία δεν περιγράφετε πολύ καλά στα σετ εκμάθησης, η εκδοχή με επιλογή πρόγνωσης υπερτερεί της εκδοχής χωρίς πρόγνωση. Αυτό υπονοεί ότι η χρήση πρόγνωσης μειώνει την ευαισθησία του κβαντιστή στις μη αντιστοιχίες μεταξύ της εκπαίδευσης και των τεστ σετ λόγο το φιλτραρίσματος της ομιλίας.

Παράμετροι	2.4kb/s	1.7kb/s
LSF's	25	21
Fourier magnitudes	8	0
Gain	8	5
Pitch and overall voicing	7	6
Bandpass voicing	4	2
Aperiodic flag	1	0
Sync bit	1	0
Total bits/frame	54	34

Πίνακας 5.2: Κατανομή των bit για τον 2.4 kb/s Federal Standard και τον 1.7 kb/s MELP κωδικοποιητή. Το μέγεθος των πλαισίων για τους δύο κωδικοποιητές είναι 22.5 ms και 20ms, αντίστοιχα.

5.4.2 Κβαντισμός των Υπόλοιπων Παραμέτρων

Ο Πίνακας 5.2 παρουσιάζει την κατανομή των bit για τον 1,7 kb/s MELP κωδικοποιητή σε σύγκριση με τον 2.4kb/s Federal Standard. Επιπροσθέτως των 4 bit που έχουμε «γλιτώσει» κατά την LSF κβάτηση και των 8 bit με τη μη μετάδοση των μεγεθών των σειρών Fourrier. «Γλιτώνουμε» 8 επιπλέον bit από τις υπόλοιπες παραμέτρους. Πρώτον, το κέρδος μεταδίδεται μία φορά ανά πλαίσιο σε σχέση με τις δύο φορές στο Federal Standard, καθώς το μέγεθος του πλαισίου τώρα είναι μικρότερο. Επίσης, βρήκαμε ότι 6 bit είναι αρκετά για να κβαντιστούν η θεμελιώδης συχνότητα και γενικά τα έμφωνα όταν δεν χρησιμοποιούνται μεγέθη Fourrier. Επιπλέον, ο αριθμός των bit που είναι απαραίτητα για πληροφορία ζωνοπερατού εμφώνου μειώνεται σε δύο, με επιλογή από ένα κατάλογο τεσσάρων δυνατών μερικών εμφώνων προτύπων. Ο απεριοδικός ενδείκτης (flag) αντικαθίσταται από μια ισάξια λειτουργική PCP (pitch contour perturbation) τεχνική, που δεν απαιτεί αποκλειστική μετάδοση.

5.5 Κωδικοποίηση Καναλιού

Προς τα εμπρός διόρθωση λαθών (FEC) κώδικες χρησιμοποιούνται για να βελτιωθεί η απόδοση στα λάθη καναλιού. Κάθε 40 ms, δύο πλαίσια που αντιστοιχούν σε δεδομένα ομαδοποιούνται και κωδικοποιούνται με ένα συνελικτικό κώδικα ρυθμού 3/5. Για να μειωθεί ο συνολικός ρυθμός μετάδοσης, το λιγότερο αντιληπτό σημαντικό τέταρτο στάδιο των LSF's αφήνεται απροστάτευτο. Προσμετρώντας ένα 4ων bit CRC που προστατεύει τα σημαντικότερα bits και μια ουρά 6 bit, ο γενικός ρυθμός μετάδοσης στο κανάλι είναι 3kb/s. Στην πλευρά του δέκτη, ένας Viterbi αποκωδικοποιητής δέχεται

τα υπό εξέταση δεδομένα από τον αποδιαμορφωτή και εκτελεί μια Μέγιστη-Likelihood αποκωδικοποίηση. Αν ο CRC εντοπίσει ένα λάθος, ένας αλγόριθμος εξάλειψης πλαισίου εξάγει συμπερασματικά τιμές για τις παραμέτρους του παρόντος πλαισίου από την προηγούμενη ιστορία.

5.6 Αποτελέσματα Υποκειμενικών Τεστ

Ο κωδικοποιητής υποβλήθηκε σε υποκειμενικό τεστ ακρόασης. Στο τεστ αυτό ο ακροατής έπρεπε να επιλέξει την καλύτερη μεταξή δύο. Για το τεστ αυτό χρησιμοποιήθηκαν 102 ζευγάρια προτάσεων, επωμένες από 10 διαφορετικούς ομιλητές, με κωδικοποιητη αναφοράς τον 2.4 kb/s Federal Standard. Το υλικό του τεστ περιείχε καθαρή ομιλία, και ομαλή και IRS φιλτραρισμένη, καθώς και διαφορετικά είδη θορύβου (γραφείου, κίνησης κτλ.). Τα ζεύγη μπήκαν σε τυχαία σειρά και παρουσιάστηκαν σε πέντε διαφορετικούς ακροατές. Γενικά ο νέος χαμηλού ρυθμού κωδικοποιητης MELP προτιμήθηκε σε σχέση με τον Federal Standard, με καθαρή προτίμηση σε πέντε από τις έξι συνθήκες του τεστ. Μόνο για ομαλή και χωρίς θόρυβο ομιλία προτιμήθηκε ο Federal Standard, λόγο της παρουσίας των μεγεθών των σειρών Fourrier.

5.7 Συμπεράσματα

Εδώ παρουσιάστηκε ένας νέος κωδικοποιητής MELP ο οποίος, μέσω βελτιώσεων του μοντέλου και της κβάντισης, υπερτερεί του Fenteral Standard σε σημαντικά χαμηλότερο ρυθμό μετάδοσης, κάνοντας τον έτσι ελκυστικό στις ασύρματες επικοινωνίες άλλα και για τις εφαρμογές χαμηλού ρυθμού δεδομένων, καθώς μας δίνει καλύτερη ποιότητα εξόδου ακόμα και σε πολύ υποβαθμισμένα κανάλια.

ΚΕΦΑΛΑΙΟ 6

6.1 Εισαγωγή

Αυτό το κεφάλαιο αναφέρεται στην υλοποίηση του κωδικοποιητή Pitch Excited LPC με τη χρήση του προγράμματος Matlab. Η υλοποίηση αυτή γίνεται με τη μορφή κώδικα και τη χρήση m files και μας προσφέρει τη δυνατότητα να μεταβάλουμε βασικές παραμέτρους έτσι ώστε να πάρουμε το βέλτιστο δυνατό αποτέλεσμα.

6.2 Υλοποίηση Κωδικοποιητή

Ο κωδικοποιητής που υλοποιούμε είναι ο Pitch Excited LPC και για την υλοποίηση του βασιστήκαμε στο θεωρητικό υπόβαθρο που περιγράφεται στα κεφάλαια 2 και 3. Η υλοποίηση πραγματοποιήθηκε σε τέσσερα m files:

lpcprj_ main.m: Η συνάρτηση αυτή είναι ο συνδετικός κρίκος μεταξύ των m file καθώς μέσα από αυτή καλούνται όλες οι συναρτήσεις που είναι απαραίτητες για την πλήρη υλοποίηση του κωδικοποιητή. Επίσης εδώ γίνεται η εξαγωγή των LP συντελεστών, η επανασύνθεση της ομιλίας και η εξαγωγή του σηματοθορυβικού λόγου.

lpcprj_levinson.m: Εδώ χρησιμοποιούμε των αλγόριθμο του Levinson για τον υπολογισμό των βέλτιστων συντελεστών γραμμικής πρόγνωσης και το ελάχιστο μέσου τετραγώνου σφάλμα πρόγνωσης.

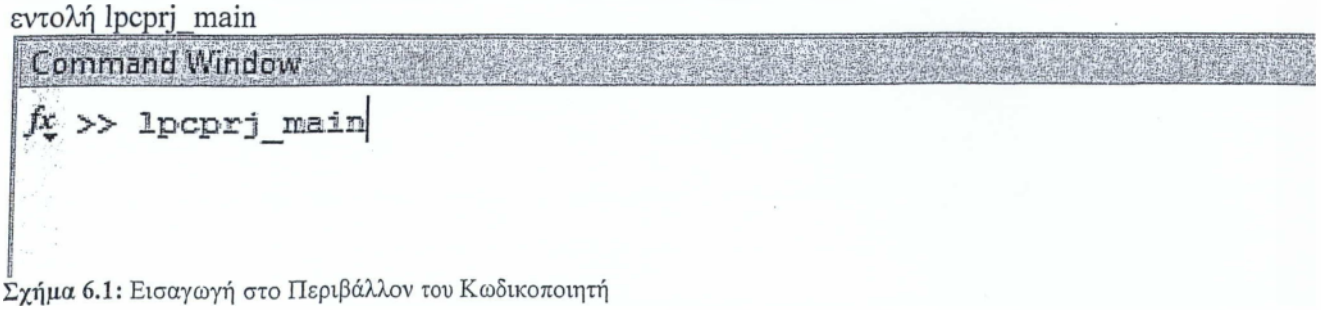
lpcprj_ pitchdetect.m: Εδώ γίνεται η εύρεση της θεμελιώδους συχνότητας κάθε πλαισίου με τη χρήση της μεθόδου της αυτοσυσχέτισης.

lpcprj_ setparams.m: Αυτή η συνάρτηση μας επιτρέπει να εισάγουμε τις παραμέτρους πάνω στις οποίες θα λειτουργήσει ο κωδικοποιητής μας μέσω ενός διαδραστικού περιβάλλοντος.

Ο πλήρης κώδικας της υλοποίησης τους φαίνεται στο παράρτημα που ακολουθεί στο τέλος του κεφαλαίου.

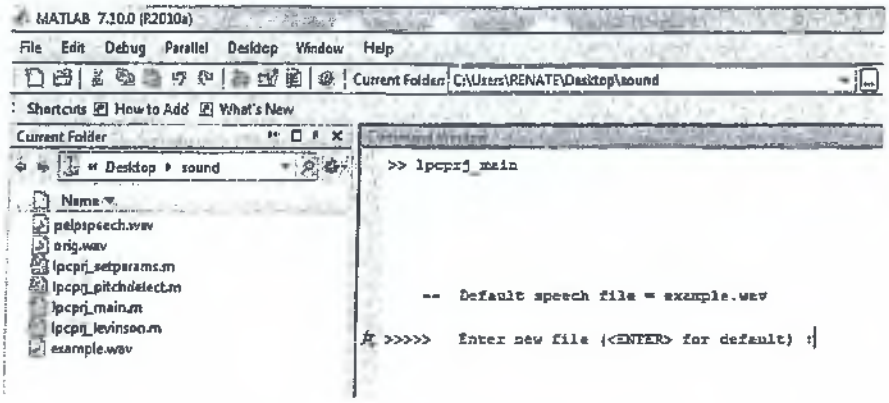
6.3 Λεπτομέρειες Χρήσης

Για να μπορέσουμε να κάνουμε χρήση του κωδικοποιητή είναι απαραίτητο να έχουμε κάνει εγκατάσταση του προγράμματος Matlab στον υπολογιστή μας. Στη συνέχεια τοποθετούμε τα m file στο φάκελο C:\Users\RENATE\Desktop\sound . Αφού ανοίξουμε το περιβάλλον του MATLAB στη γραμμή εντολών δίνουμε την εντολή lpcprj_main



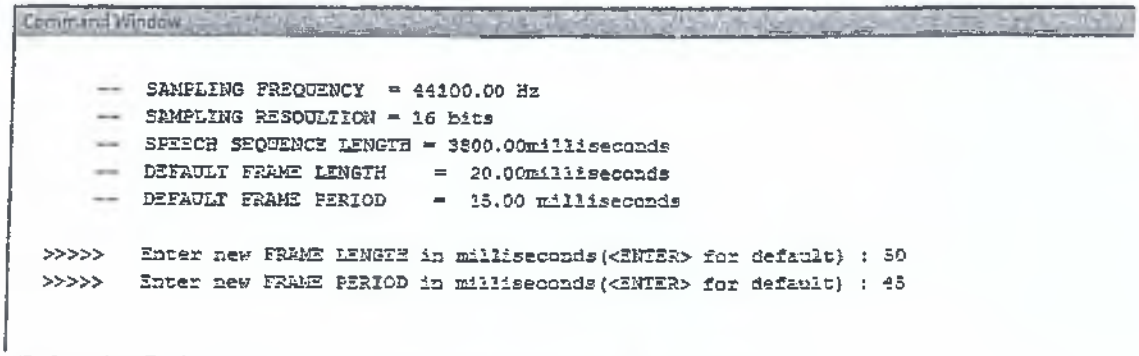
Σχήμα 6.1: Εισαγωγή στο Περιβάλλον του Κωδικοποιητή

Με την εκτέλεση αυτής της εντολής θα εμφανιστεί ένα πλαίσιο στο οποίο θα μας ζητείτε να εισάγουμε ένα αρχείο.wav ή να πατώντας enter να επιλέξουμε το προεπιλεγμένο (τα αρχεία.wavπρέπει να βρίσκονται και αυτά στον κατάλογο C:\Users\RENATE\Desktop\sound).



Σχήμα 6.2: Επιλογή του Αρχείου Ομιλίας

Έχοντας επιλέξει το example.wav που μας ενδιαφέρει το πρόγραμμα μας παραθέτει ορισμένες πληροφορίες που σχετίζονται με το αρχείο μας, όπως η συχνότητα με την οποία έχει δειγματολυπτηθεί, τον αριθμό των bit με τον οποίο έχει κωδικοποιηθεί και το συνολικό μήκος του αρχείου μας σε ms. Επίσης μας προτείνει ένα μέγεθος πλαισίου και την περίοδο με την οποία θέλουμε να επαναλαμβάνεται το πλαίσιο (Εδώ πρέπει να πούμε ότι η περίοδος δεν μπορεί να είναι μεγαλύτερη του μεγέθους του πλαισίου), δίνοντας μας όμως τη δυνατότητα να τα αλλάξουμε.



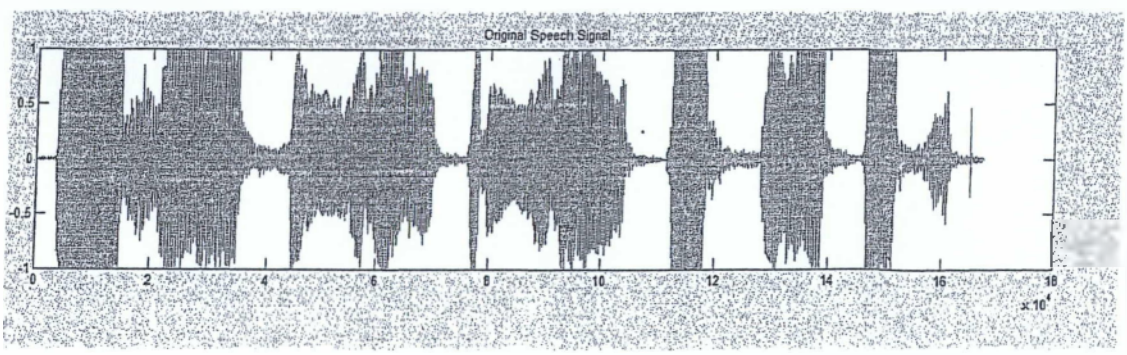
Σχήμα 6.3: Επιλογή του Μεγέθους Πλαισίου και της Περιόδου πλαισίου

Στη συνέχεια επιλέγουμε το παράθυρο που θα χρησιμοποιηθεί κατά τη διαδικασία της LP ανάλυσης καθώς και την τάξη του AR μοντέλου.

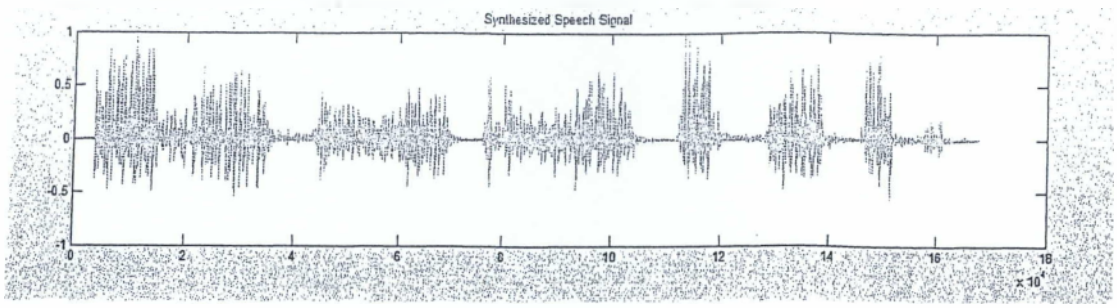
```
      --  DEFAULT FRAME WINDOW    = hamming

      --  0 = BOXCAR
      --  1 = BARTLETT
      --  2 = HAMMING
      --  3 = HANNING

  >>>>>   Enter new FRAME WINDOW by number (<ENTER> for default) : 2




      --  DEFAULT AR MODEL ORDER = 16

  >>>>>   Enter a new AR MODEL ORDER (<ENTER> for default) : 16
```

Σχήμα 6.4: Επιλογή Παραθύρου και AR Μοντέλου.

Από αυτή τη στιγμή και πέρα αρχίζει η όλη επεξεργασία και κωδικοποίηση του σήματος. Με το πέρας της διαδικασίας ο κωδικοποιητής μας έχει ως έξοδο το κωδικοποιημένο σήμα το οποίο και σώζεται στο αρχείο pelpspeech.wav, το σηματοθορυβικό λόγο σε dB και τη γραφική αναπαράσταση του αυθεντικού σήματος ομιλίας και του κωδικοποιημένου σήματος.



Σχήμα 6.5: Original speech signal



Σχήμα 6.6: Synthesized speech signal

6.4 Αποτελέσματα Συμπεράσματα

Για να μπορέσουμε να έχουμε μια πιο πλήρη εικόνα για τον κωδικοποιητή μας και για να μπορέσουμε να αξιολογήσουμε την απόδοση του, χρησιμοποιούμε τις αντικειμενικές τεχνικές αξιολόγησης και συγκεκριμένα το σηματοθορυβικό λόγο (SNR). Για το λόγο αυτό υποβάλαμε τον κωδικοποιητή μας στην εξής διαδικασία: κρατώντας σταθερή την επιλογή του παραθύρου (Hamming) και την τάξη του AR μοντέλου (16),μεταβάλουμε ταυτόχρονα το μήκος του πλαίσιο και την επανάληψη του πλαισίου κατά 5 ms, ξεκινώντας από αρχικές τιμές 10 ms και 5 ms αντίστοιχα, και καταλήγοντας να έχουμε ένα πλαίσιο μεγέθους 50 ms και επανάληψη πλαισίου κάθε 45 ms. Τα αποτελέσματα φαίνονται στον παρακάτω Πίνακα 6.1

Δοκιμές	Frame Length	Frame Period	SNR (dB)	Processed frames
1	10	5	60.46	759
2	15	10	55.02	380
3	20	15	49.40	254
4	25	20	Inf	190
5	30	25	66.87	152
6	35	30	51.60	127
7	40	35	47.39	109
8	45	40	45.36	95
9	50	45	43.31	85

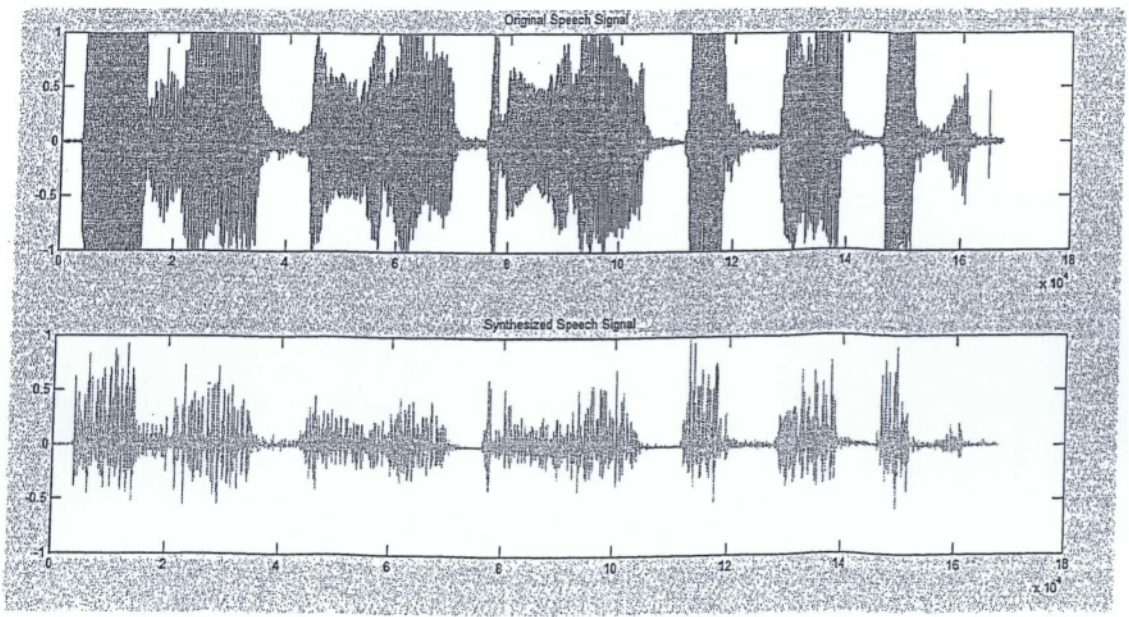
Πίνακας 6.1: Αποτελέσματα Κωδικοποιητή

Επιλέξαμε την παραπάνω διαδικάσία διότι είδαμε ότι αν κρατάμε σταθερά το μήκος του πλαισίου αλλά και την περίοδο του πλαισίου και μεταβάλλουμε τον τύπο του παραθύρου και την τάξη του AR μοντέλου , ο σηματοθορυβικός λόγος και ο αριθμός των πλαισίων προς επεξεργασία παραμένει σταθερός.Ωστόσο βλέπουμε ότι η ποιότητα της παραγόμενης ομιλίας έχει μια διαφορά. Αυτό βέβαια μπορεί να τεκμηριωθεί υποβάλλοντας τον κωδικοποιητή σε υποκειμενικές τεχνικές αξιολόγησης (MOS, DAM, DRT Κεφάλαιο 1). Όμως δεν πρέπει να ξεχνάμε πως ο κωδικοποιητής ανήκει στην κατηγορία των αλγορίθμων με απώλειες με αποτέλεσμα η ομιλία που παράγεται να είναι συνθετική.

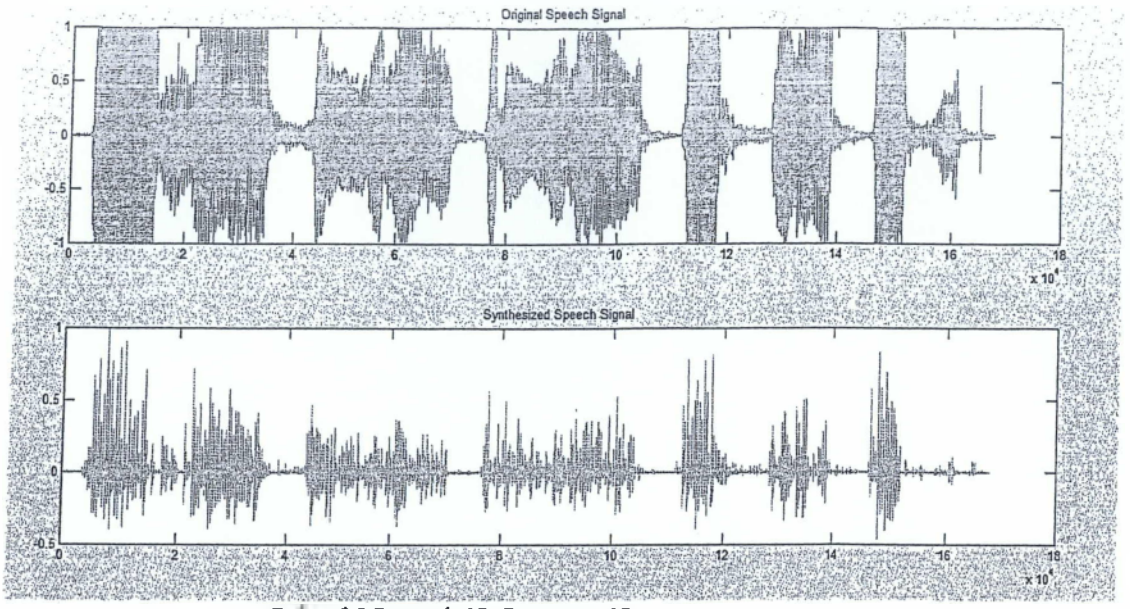
Αν παρατηρήσουμε τα αποτελέσματα μας βλέπουμε ότι για μικρά μεγέθη

πλαισίου και μικρή περίοδο επανάληψης του πλαισίου ο σηματοθορυβικός λόγος μας δίνει πολύ καλά αποτελέσματα με προφανές κόστος το χρόνο επεξεργασίας καθώς ο αριθμός των πλαισίων που πρέπει να επεξεργασθούν είναι αυξημένος. Ωστόσο , αν χρησιμοποιήσουμε μεγάλα μεγέθη πλαισίου και μεγάλη περίοδο επενάληψης του πλαισίου , έχουμε μεν εξοικονόμηση στο χρόνο επεξεργασίας όμως ο σηματοθορυβικός λόγος ειώνεται εκτός δύο περιπλωσεων όπως είναι φανερό και στον πίνακα μας με συνέπεια την ενίσχυση του θορύβου στο σήμα μας.

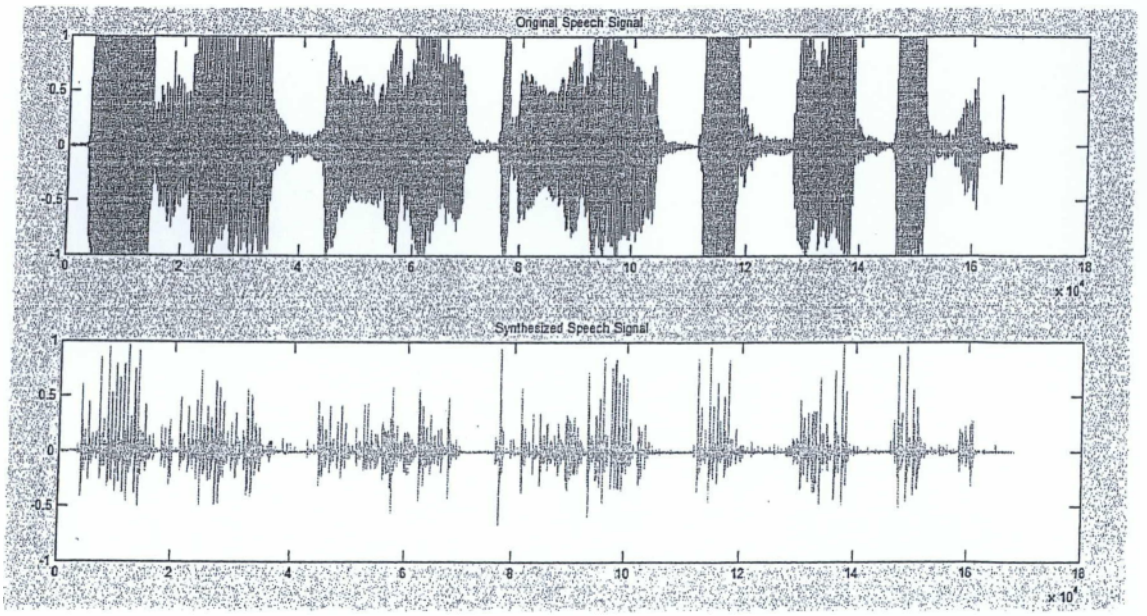
Αν ψάξουμε να βρούμε τη χρυσή τομή μεταξύ του χρόνου επεξεργασίας και του σηματοθορυβικού λόγου θα πρέπει να βασιστούμε στην υποκειμενική μας αντίληψη, καθώς οι δοκιμές 7 και 9 μας έδωσαν λίγα πλαίσια προς επεξεργασία και σχετικά καλό σηματοθορυβικό λόγο 47,39 και 44 dB και 109 και 85 frames αντίστοιχα. Όμως η ποιότητα του σήματος εξόδου ήταν ιδιαίτερα υποβαθμισμένη σε σχέση με την ποιότητα του σήματος εξόδου κατά τη δοκιμή 3 παρόλο που ο σηματοθορυβικός λόγος στη δοκιμή 3 υπολείπεται των σηματοθορθβικών λόγων των δοκιμών 7 και 9.



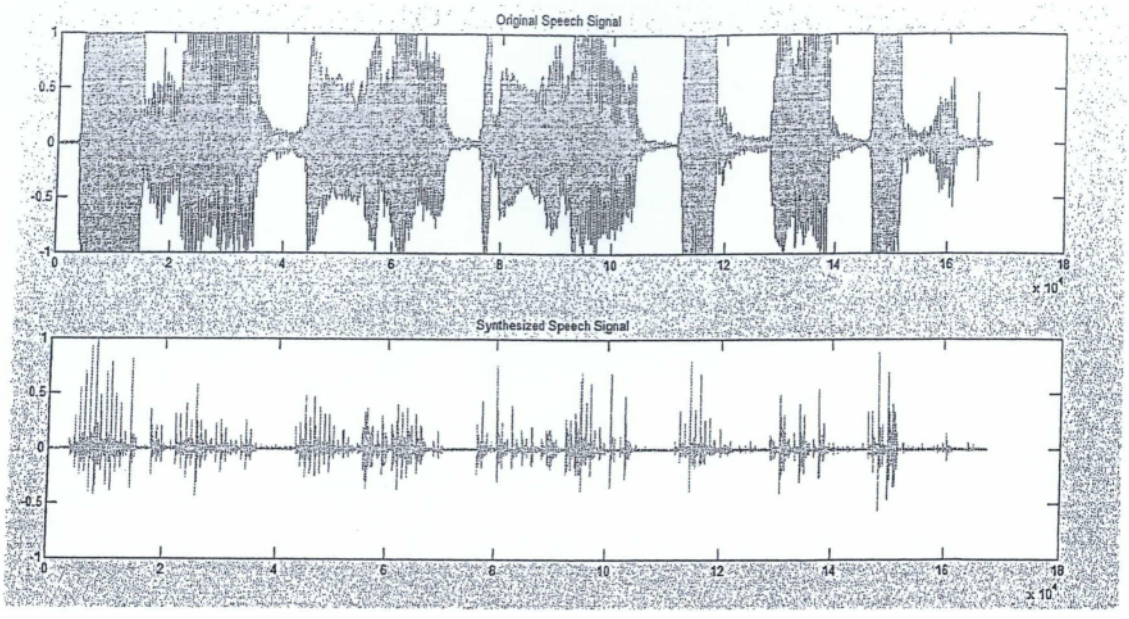
Σχήμα 6.7 Frame I. 10 -Frame p. 5



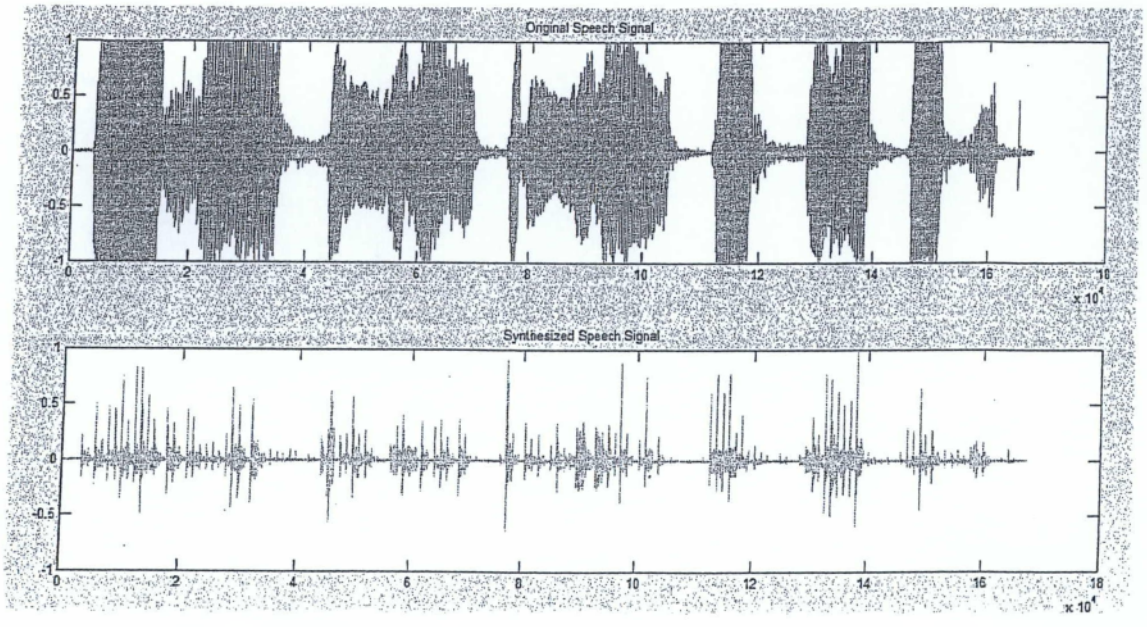
Σχήμα 6.8 Frame l. 15 -Frame p. 10



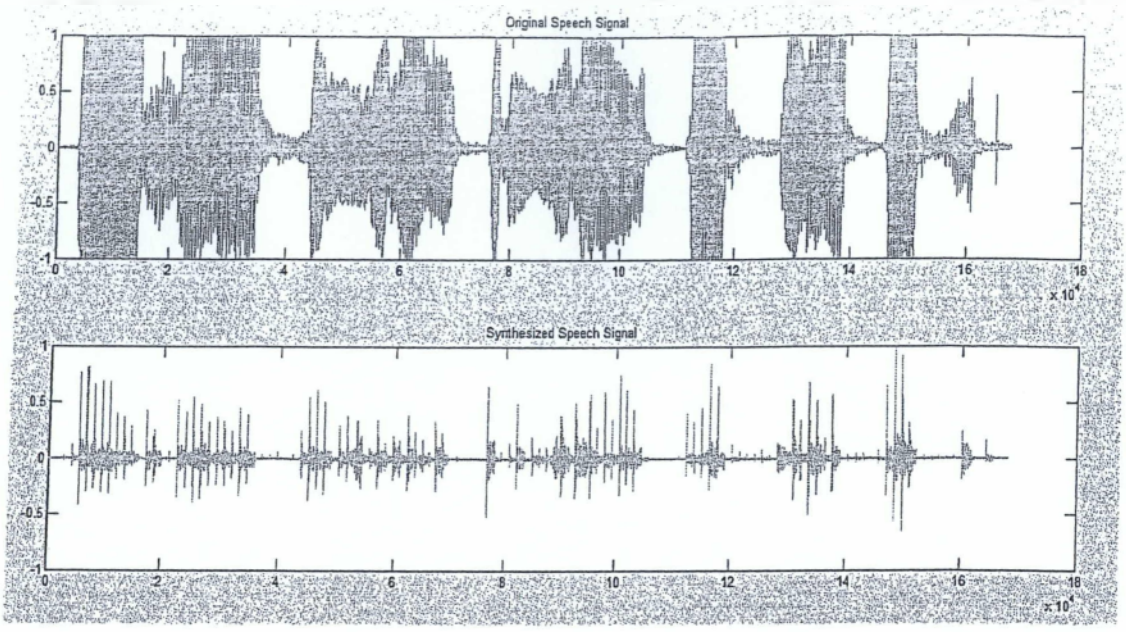
Σχήμα 6.9 Frame l. 20 -Frame p. 15



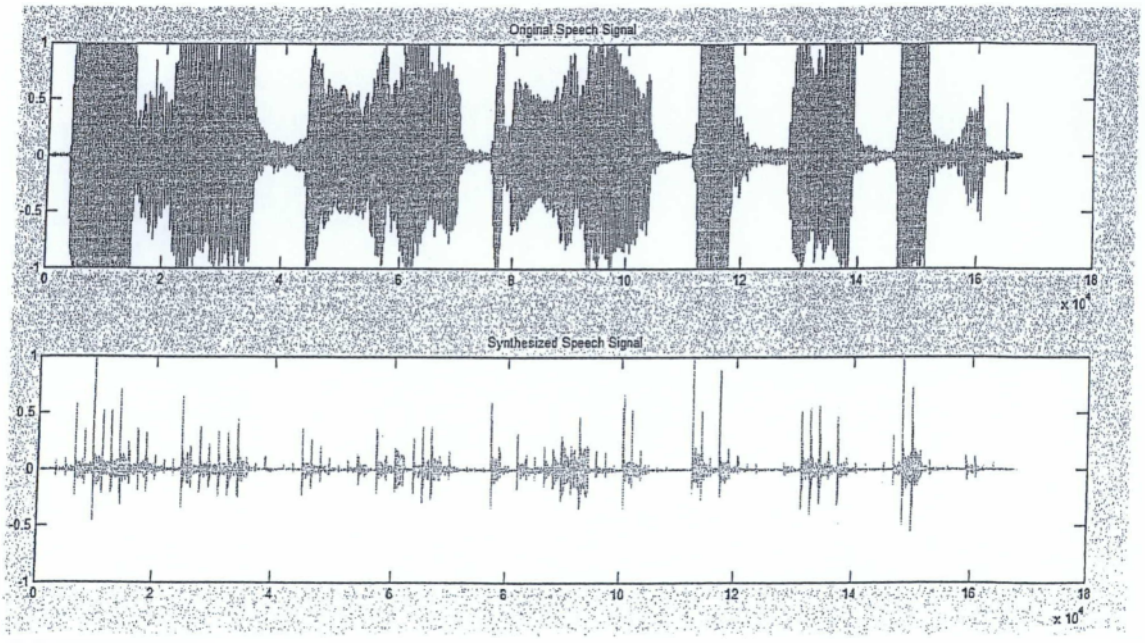
Σχήμα 6.10 Frame I.25 -Frame p. 20



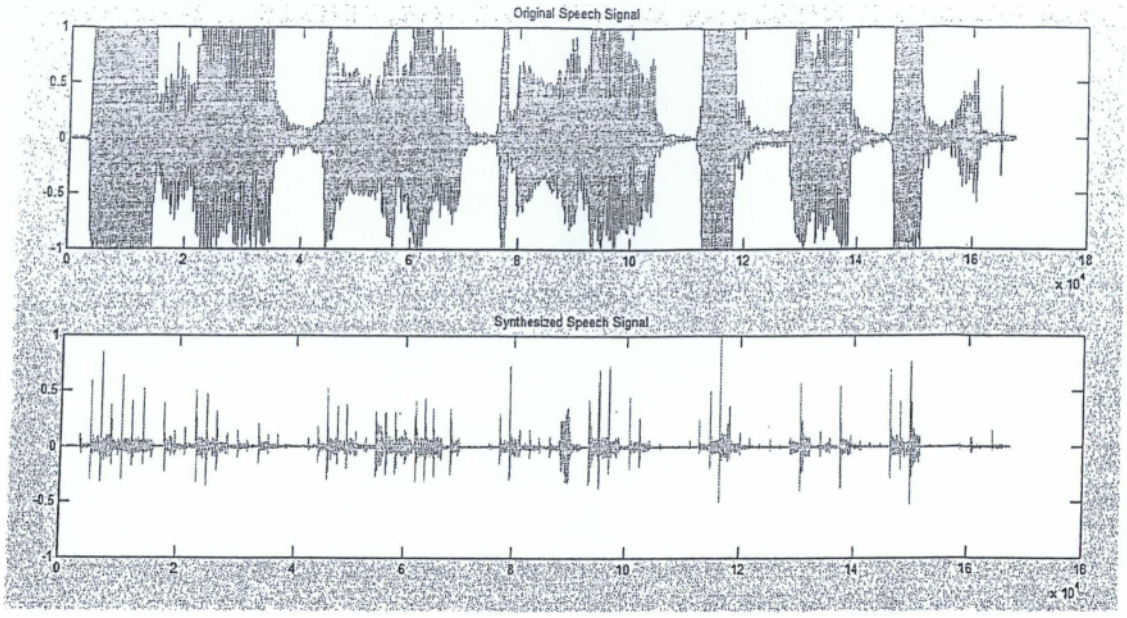
Σχήμα 6. 11 Frame I. 30 -Frame p. 25



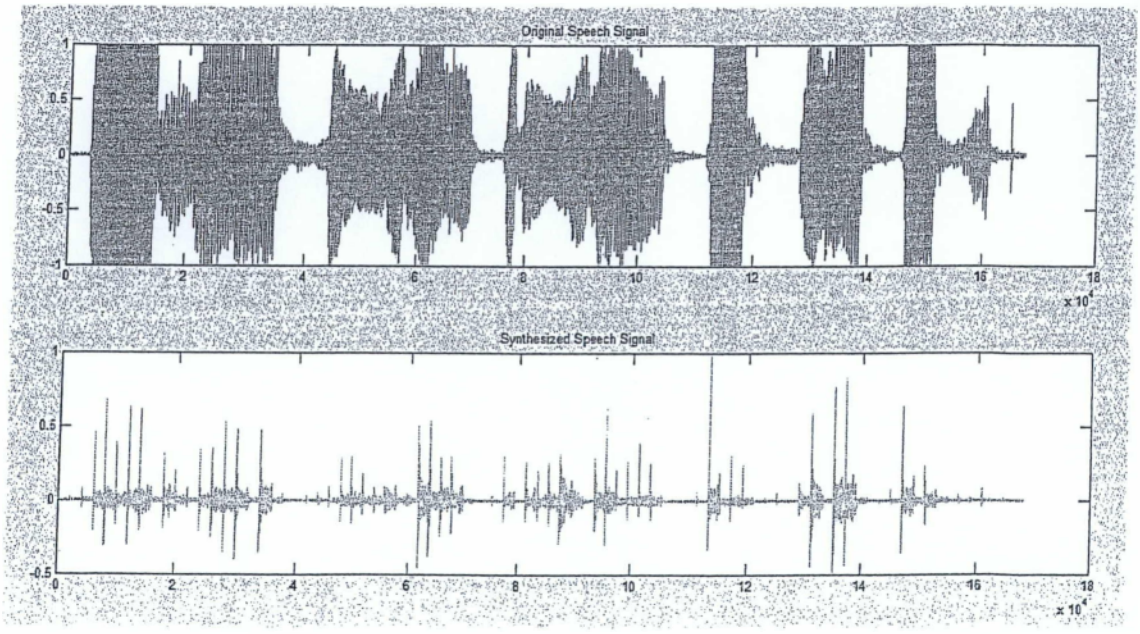
Σχήμα 6.12 Frame l. 35 -Frame p. 30



Σχήμα 6.13 Frame l. 40 -Frame p. 35

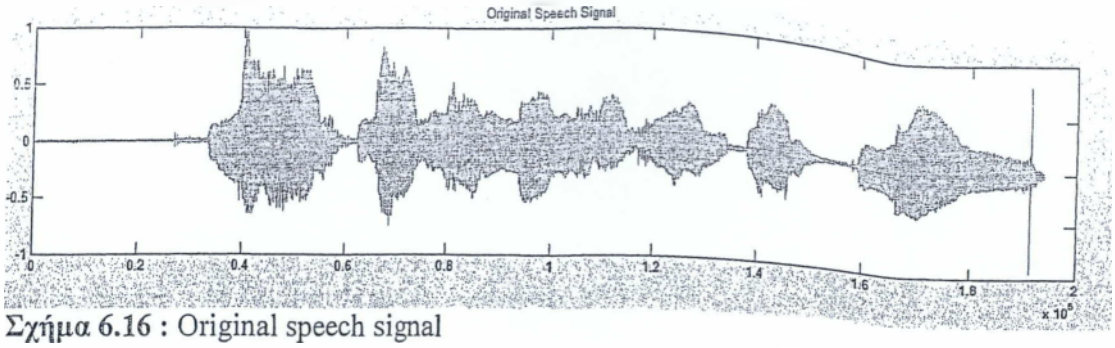


Σχήμα 6.14 Frame l.45 -Frame p. 40

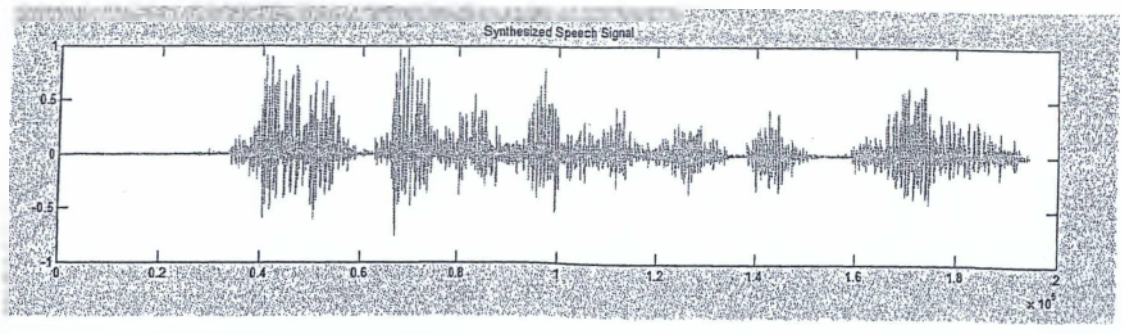


Σχήμα 6.15 Frame l. 50 -Frame p. 45

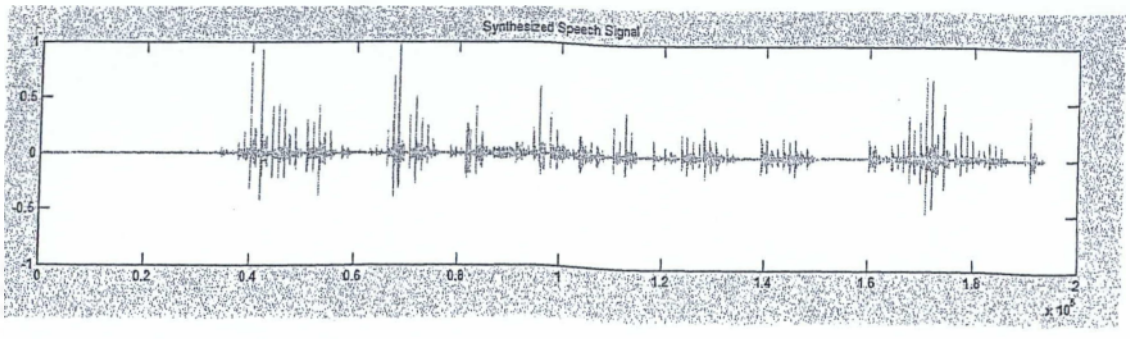
Ακολουθούν οι εικόνες που μας δείχνουν την κωδικοποίηση της γυναικείας φωνής.



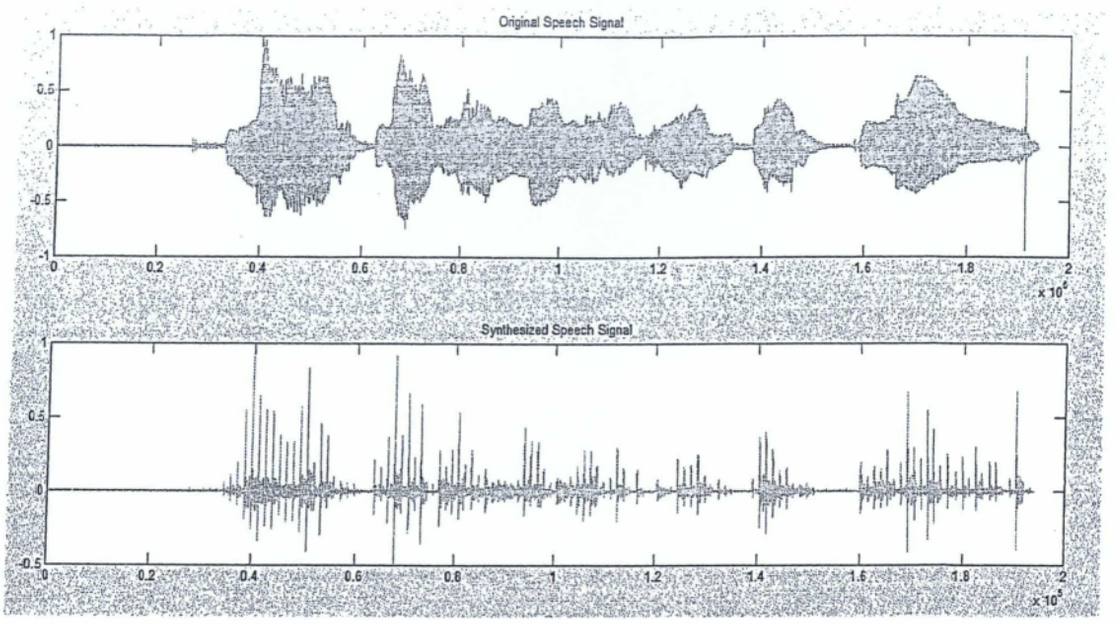
Σχήμα 6.16 : Original speech signal



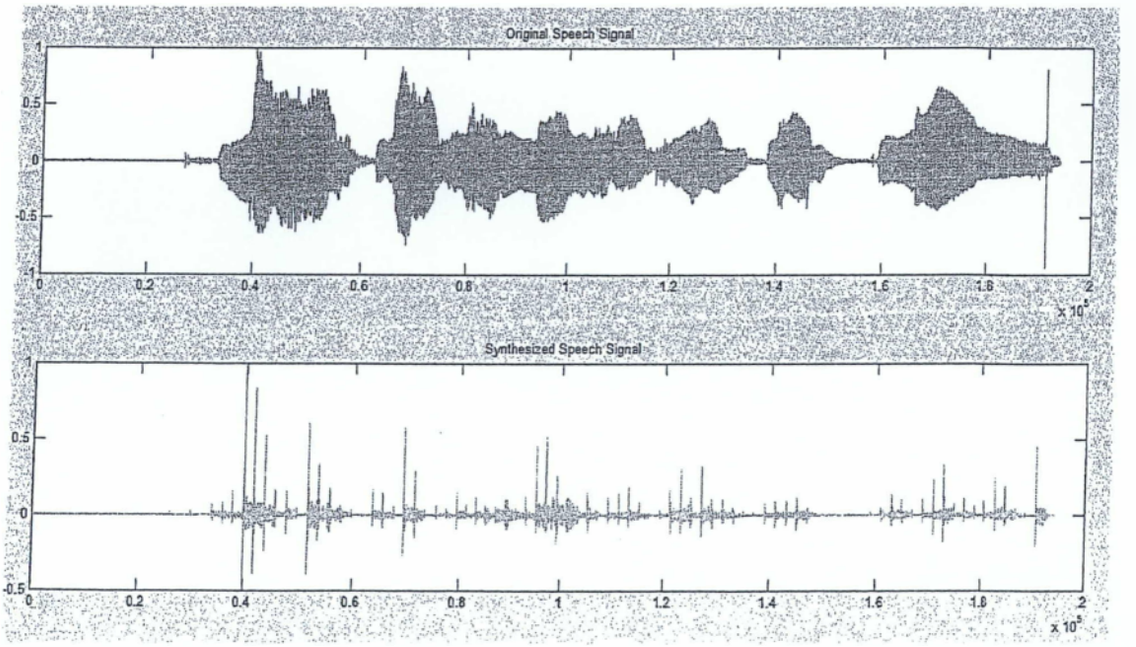
Σχήμα 6.17 : Synthesized speech signal



Σχήμα 6.18: Frame l. 30 frame p. 25



Σχήμα 6.19: Frame l. 35 frame p. 30



Σχήμα 6.20: Frame l. 50 frame p. 45

Δοκιμές	Frame Length	Frame Period	SNR (dB)	Processed frames
1	10	5	58.95	879
2	15	10	inf	440
3	20	15	50.37	294
4	25	20	Inf	220
5	30	25	66.87	176
6	35	30	52.87	147
7	40	35	51.71	126
8	45	40	51.82	110
9	50	45	51.95	98

Σχήμα 6.21: Εισαγωγή στο Περιβάλλον του Κωδικοποιητή 2

ΠΑΡΑΡΤΗΜΑ

lpcprj_main.m

```
clear all;
debug = 0;
default_speechfile  = 'five2.wav';
default_frametime   = 20.0;
default_frameperiod = 15.0;
default_framewindow = 'hamming';
default_modelorder  = 16;

[speechfile,speechlength,speechrate,speechres,framelength,framedelta,..
.
   framewindow,modelorder,prefilterbvec,prefilteravec] =...
   lpcprj_setparams(default_speechfile,default_frametime,...
   default_frameperiod,default_framewindow,default_modelorder);

totalframes = ceil(speechlength / framedelta);


switch (framewindow)
case {'boxcar'},
   windowfunc = boxcar(framelength);
case {'bartlett'},
   windowfunc = bartlett(framelength);
case {'hamming'},
   windowfunc = hamming(framelength);
case {'hanning'},
   windowfunc = hanning(framelength);
end


fprintf('\n\n\n\n\n');
speechindex1 = 1;
for framenum = 1:totalframes

   fprintf('Processing frame %5d of %5d\n',framenum,totalframes);

   speechindex2 = speechindex1 + framelength - 1;
   if (speechindex2 > speechlength)
      if (speechindex1 > speechlength)
         speechframe = zeros(framelength,1);
      else
         tempframe   = wavread(speechfile,[speechindex1 speechlength]);
         padlength   = framelength - ((speechlength-speechindex1)+1);
         speechframe = [tempframe(:,1) ; zeros(padlength,1)];
```

```
      end
   else
      tempframe   = wavread(speechfile,[speechindex1 speechindex2]);
      speechframe = tempframe(:,1);
   end

   coeffsframe = filter(prefilterbvec,prefilteravec,speechframe);
   coeffsframe = coeffsframe .* windowfunc;
   rxx         = xcorr(coeffsframe);
   anxt        = inf;
   rhonxt      = inf;
   for order = 1:modelorder
      a     = anxt;
      rho   = rhonxt;
      clear anxt;
      clear rhonxt;
      [anxt,rhonxt,k,status] = lpcprj_levinson(a,rho,rxx);
      clear a;
      clear rho;
   end
   arcoeffs = [1 ; anxt];

   sourceframe  = filter(arcoeffs,1,speechframe);
   sourceenergy = sum(sourceframe .* sourceframe);
   period       = lpcprj_pitchdetect(sourceframe);

   if (period == 0);
      amplitude  = sqrt(sourceenergy/framelength);
      excitation = randn(framelength,1) * amplitude;
   else
      amplitude  = sqrt((.15*sourceenergy)/framelength);
      excitation = randn(framelength,1) * amplitude;
      spikecount = floor(framelength / period) + 1;
      amplitude  = sqrt((.85*sourceenergy)/spikecount);
      index = 1;
      while (index <= framelength)
         excitation(index) = excitation(index) + amplitude;
         index = index + period;
      end
   end

   synthframe   = filter(1,arcoeffs,excitation); speechindex2 =
   speechindex1 + framedelta - 1;
   speechrep(speechindex1:speechindex2) = synthframe(1:framedelta);
   speechindex1 = speechindex1 + framedelta;

if (debug == 1)
   if (period ~= 0)
      figure(1);
      subplot(3,1,1);
      plot(speechframe,'b');
```

```
        axis tight; hold on;
        plot(synthframe,'r');
        axis tight;
        hold off;

        subplot(3,1,2);
        plot(sourceframe,'b');
        axis tight;
        hold on;
        plot(excitation,'r');
        axis tight;
        hold off;

        z = xcorr(speechframe);

        subplot(3,1,3);
        plot(xcorr(excitation),'r');
        axis tight;
        hold on;
        plot(xcorr(sourceframe),'b');
        axis tight;
        hold off;

        period
        xxx = input('Press a key...');
    end
end

end
[b a]      = butter(15,.3);
speechrep  = speechrep/max(abs(speechrep));
filtspeech = filter(b,a,speechrep);
filtspeech = filtspeech/max(abs(filtspeech));
fslength=length(filtspeech);
if length(fslength)<length(speechlength)
   for j=1:length(fslength)
      N=N+speechlength(j)^2;
      D=D+(speechlength(j)-fslength(j))^2;
   end
else
   for j=1:length(speechlength)
      N=N+speechlength(j)^2;
      D=D+(speechlength(j)-fslength(j))^2;

   end
end

S=10*log10(N/D);

fprintf('Signal to noise ratio is %.2fdb\n',S);
sound(filtspeech,speechrate,speechres);
wavwrite(filtspeech,speechrate,speechres,'pelpspeech.wav');
subplot(2,1,2);
```

```
plot(filtspeech,'g');
title('Synthesized Speech Signal');
```

lpcprj_setparams.m

```
function
[speechfile,speechlength,speechrate,speechres,framelength,framedelta,..
.
   framewindow,modelorder,prefilterbvec,prefilteravec] =...
   lpcprj_setparams(default_speechfile,default_frametime,...
   default_frameperiod,default_framewindow,default_modelorder);


status = -1;
while (status == -1)
   fprintf('\n\n\n\n\n');
   fprintf('   --  Default speech file = %s\n',default_speechfile);
   fprintf('\n');
   speechfile = input('>>>>> Enter new file (<ENTER> for default) :
','s');
   if isempty(speechfile)
      speechfile = default_speechfile;
   end
   status = fopen(speechfile,'r');
   if (status == -1)
      fprintf('\n');
      fprintf('   --  ERROR ----------------------------------------
--------\n');
      fprintf('\n');
      fprintf('   --  %s cannot be opened.  It may be misspelled,
\n',speechfile);
      fprintf('   --  may not exist, or may be locked by another
application.\n');
      fprintf('\n');
      fprintf('   --  ERROR ----------------------------------------
--------\n');
      pause(2);
   end
end
fclose(status);
[dummy,speechrate,speechres] = wavread(speechfile);
wavwrite(dummy,speechrate,speechres,'orig.wav');
[originalsignal,Fsoriginal,Originalbits]=wavread('c:\matlabr11\work\ori
g.wav');
wavplay(originalsignal,Fsoriginal)
subplot(2,1,1);
plot(originalsignal);
title('Original Speech Signal');
templength = wavread(speechfile,'size');
speechlength = templength(1);
if (templength(2) > 1)
```

```
      fprintf('\n');
      fprintf('   --  WARNING ----------------------------------------
   -----\n');
      fprintf('\n');
      fprintf('   --  This file contains multi-channel data.  Only
   channel\n');
      fprintf('   --  one will be processed.
   \n');
      fprintf('\n');
      fprintf('   --  WARNING ----------------------------------------
   -----\n');
      pause(2);
   end
   speechtime = (speechlength / speechrate) * 1000;


   status = -1;
   while (status == -1)
      status = 0;
      fprintf('\n\n\n\n\n');
      fprintf('   --  SAMPLING FREQUENCY     = %6.2f Hz\n',speechrate);
      fprintf('   --  SAMPLING RESOULTION    = %d bits\n',speechres);
      fprintf('   --  SPEECH SEQUENCE LENGTH = %6.2f
   milliseconds\n',speechtime);
      fprintf('   --  DEFAULT FRAME LENGTH   = %6.2f
   milliseconds\n',default_frametime);
      fprintf('   --  DEFAULT FRAME PERIOD   = %6.2f
   milliseconds\n',default_frameperiod);
      fprintf('\n');
      frametime   = input('>>>>>  Enter new FRAME LENGTH in milliseconds
   (<ENTER> for default) : ');
      if isempty(frametime)
         frametime = default_frametime;
      end
      frameperiod = input('>>>>>  Enter new FRAME PERIOD in milliseconds
   (<ENTER> for default) : ');
      if isempty(frameperiod)
         frameperiod = default_frameperiod;
      end
      if ((frametime < 1)|(frameperiod < 1)|(frametime >
   speechtime)|(frameperiod > frametime))
         status = -1;
         fprintf('\n');
         fprintf('   --  ERROR ------------------------------------------
   --------\n');
         fprintf('\n');
         fprintf('   --  Minimum FRAME LENGTH = 1 millisecond.
   \n');
         fprintf('   --  Maximum FRAME LENGTH = SPEECH SEQUENCE
   LENGTH\n');
         fprintf('   --  Minimum FRAME PERIOD = 1 millisecond.
   \n');
```

```
        fprintf('   --  Maximum FRAME PERIOD = FRAME LENGTH
 \n');
        fprintf('\n');
        fprintf('   --  ERROR ----------------------------------------
 --------\n');
       pause(4);
     end
  end
  framelength = ceil((frametime/1000) * speechrate);
  framedelta  = ceil((frameperiod/1000) * speechrate);

  status = -1;
  while (status == -1)
     status = 0;
     fprintf('\n\n\n\n\n\n');
     fprintf('   --  DEFAULT FRAME WINDOW    = %s\n',default_framewindow);
     fprintf('\n');
     fprintf('   --      0 = BOXCAR\n');
     fprintf('   --      1 = BARTLETT\n');
     fprintf('   --      2 = HAMMING\n');
     fprintf('   --      3 = HANNING\n');
     fprintf('\n');
     windownum   = input('>>>>>  Enter new FRAME WINDOW by number
  (<ENTER> for default) : ');
     if isempty(windownum)
        framewindow = default_framewindow;
     else
        switch (windownum)
        case {0},
           framewindow = 'boxcar';
        case {1},
           framewindow = 'bartlett';
        case {2},
           framewindow = 'hamming';
        case {3}
           framewindow = 'hanning';
        otherwise, status
           = -1;
           fprintf('\n');
           fprintf('   --  ERROR ----------------------------------------
 ----------\n');
           fprintf('\n');
           fprintf('   --  Input does not match any available numeric
 selection.    \n');
           fprintf('\n');
           fprintf('   --  ERROR ----------------------------------------
 ----------\n');
           pause(1);
       end
    end
 end
```

```
   status = -1;
   while (status == -1)
      status = 0;
      fprintf('\n\n\n\n\n');
      fprintf('   --  DEFAULT AR MODEL ORDER = %d\n',default_modelorder);
      fprintf('\n');
      modelorder   = input('>>>>>  Enter a new AR MODEL ORDER (<ENTER> for
   default) : ');
      if isempty(modelorder)
         modelorder = default_modelorder;
      end
      if ((modelorder < 1) | (modelorder ~= floor(modelorder)))
         status = -1;
         fprintf('\n');
         fprintf('   --  ERROR ----------------------------------------
   --------\n');
         fprintf('\n');
         fprintf('   --  AR model order must be an integer.
   \n');
         fprintf('   --  AR model order must be greater than 0.
   \n');
         fprintf('\n');
         fprintf('   --  ERROR ----------------------------------------
   --------\n');
        pause(2);
      end
   end

   prefilterbvec = [1 ; 1/2];
   prefilteravec = 1;
```

lpcprj_pitchdetect.m

```
   y = speechframe;
   n = length(speechframe);

   rxxslope   = .125;
   datalength = length(y);
   rxxfix     = [(datalength:-1:1) (2:1:datalength)]' * rxxslope;
   rxx        = xcorr(y) .* rxxfix;
   rxx        = rxx .* hamming(length(rxx));
   rxxfix     = rxxfix + (1-rxxslope);
   temp       = ([0 ; rxx(n:length(rxx)) ; 0]);
   corrlength = length(temp) - 2;

   rxxpeakloc = 0;
   rxxpeakval = -inf;

   for index = 6:corrlength;
      prev    = temp(index);
      curr    = temp(index + 1);
      next    = temp(index + 2);
      if ((curr > prev) & (curr > next) & (curr > rxxpeakval))
         rxxpeakval = curr;
```

```
      rxxpeakloc = index;
   end
end

if ((rxxpeakval <= (0.3*rxx(n))) | (rxxpeakloc == 0))
   period = 0;
else
   period = rxxpeakloc - 1;
end;
```

lpcprj_levinson.m

```
% lpcprj_levinson.m performs one iteration of the levinson
% algorithm, an algorithm which calculates optimal linear
% prediction coefficients and minimum mean-squared prediction
% error for a data sequence using previous parameter values
% and the sequence's autocorrelation values.  It also provides
% for calculation of intial coefficient and error values.
%
% USAGE:  [anxt,rhonxt,k,status] = lpcprj_levinson(a,rho,rxx);
%
% INPUTS:  a       a COLUMN vector containing LP coefficients,
%                  with the current order determined by the
%                  number of coefficients.  If the initial
%                  values are being calculated, a must be a
%                  single-element vector with one value --
%                  infinity ("inf" in MATLAB).
%          rho     a single scalar parameter containing the
%                  minimum mean-squared error value.  If the
%                  initial values are being calculated, rho
%                  must be loaded with the value infinity.
%          rxx     a COLUMN vector of autocorrelation values for
%                  the sequence whose LP coefficients we wish
%                  to calculate.  rxx must contain an ODD
%                  number of values.  Assuming rxx consists of
%                  (2k + 1) elements, then rxx(1) equals the
%                  autocorrelation at lag -k.  rxx(2k+1) equals
%                  the autocorrelation at lag +k.  rxx(k+1)
%                  equals the autocorrelation at lag 0.
%
% OUTPUTS: anxt    a COLUMN vector containing LP coefficients
%                  for the next value of model order.  anxt
%                  generally contains one more element than
%                  the number of elements in a.  Exception:
%                  if initial values are being calculated,
%                  a contains one value (infinity) and anxt
%                  contains one value (coefficient for model
%                  order 1).
%          rhonxt  a single scalar parameter containing the
%                  next minimum mean-squared error value.
%          k       the k value used in calculating the "next"
%                  parameter values.  k obeys the following
%                  relationship: rhonxt = rho * (1 - (abs(k))^2);
%                  If initial values are being calculated, then
%                  k returns infinity (it's not used in the
%                  initialization step).
```

```
%           status  status is an integer indicating the
%                   result of calling levinson.  Each bit in
%                   status indicates the occurrence of a
%                   particular error.  If no errors occur,
%                   status = 0.  Otherwise...
%
%                   status =  1 : a has more than 1 column
%                   status =  2 : rho is not a scalar
%                   status =  4 : rxx has more than 1 column
%                   status =  8 : rxx has an even number of rows
%                   status = 16 : not enough rxx elements to
%                                 perform reqested calculation
%
%                   If multiple errors occur, then status will
%                   reflect the sum of the relevant error values.

function [anxt,rhonxt,k,status] = lpcprj_levinson(a,rho,rxx);

% Make sure that the input arguments are valid numeric
% values.  Multiplication by 1 serves as an effective test.

a   = a * 1;
rho = rho * 1;
rxx = rxx * 1;

status = 0;

[rows,cols] = size(a);              % column test
alength     = rows;
if (cols ~= 1)
   status = status + 1;
end

[rows,cols] = size(rho);            % scalar test
if ((cols ~= 1) | (rows ~= 1))
   status = status + 2;
end

[rows,cols] = size(rxx);            % rxx test
rxxlength   = rows;
rxxoffset   = ceil(rows/2);
if (cols ~= 1)
   status = status + 4;
end
if ((rows/2) == floor(rows/2))
   status = status + 8;
end

% Start of Iteration

if (status == 0)
   if ((alength == 1) & (a == inf) & (rho == inf))
      % **** Calculate initial values ****
      if (rxxlength >= 3)
         anxt      = conj(-(rxx(1+rxxoffset)/rxx(0+rxxoffset)));
         rhonxt    = (conj(anxt)*rxx(-1+rxxoffset)) + rxx(0+rxxoffset);
         status    = status + 0;
```

```
            k          = inf;
         else
            status     = status + 16;
         end
      else
         % ***** Perform a single iteration *****
         if (rxxlength >= ((2 * alength) + 3))
            p          = alength;
            predelta   = rxx(p + 1 + rxxoffset)...
                         + (conj(a(1:p))' * rxx((p:-1:1) + rxxoffset));
            delta      = conj(predelta);
            k          = -delta/rho;
            anxt       = [a ; 0] + (k * [conj(a(p:-1:1)) ; 1]);
            rhonxt     = rho * (1 - (abs(k))^2);
            status     = status + 0;
         else
            status     = status + 16;
         end
      end
   end
```

ΒΙΒΛΙΟΓΡΑΦΕΙΑ

1. Barnewell Thomas P III., Kambiz NAyebi, and Craig H. Richardon, "Speech Coding: A Computer Laboratory Textbook", *John Wiley &Sons, .Inc., 1996.*

2. Γουμενίδης Θεόδωρος, "Κωδικοποίηση Φωνής – Κωδικοποιητές Κυματομορφής", Φεβρουάριος 2004.

3. Jeremy Bradbury, "Linear Predictive Coding", December 2000.

4. NTT DoCoMo, "Proposed Contents of Operational Handbook", Asia – Pacific Telecommynity (The 3rd Meeting of the APT IMT-2000 Forum), 2-3 September 2002, Busan, Republic of Corea.

5. Jason Woodard, "Speech Coding", http://www.ecs.soton.ac.uk/~jpw/index.html

6. Niranjan Dhanakoti, "Speech Signal Processing", project report 2002.

7. Bryan Douglas, "Voice Encoding Methods for Digital Wireless Communications Systems", Fall 1997.

8. Eddie L. T. Choy, "Waveform Interpolation Speech Coder at 4 kb/s", August 1998

9. Susanna Varho, "New Linear Predictive Methods for Digital Speech Processing", 2001

10. Nadim Batri, "Robust Spectral Parameter Coding in Speech Processing", May 1998

11. Alan McCree and Jan Carlos De Martin, "A 1.7 Kb/s Melp Coder with Improved Analysis and Quantization", DSPS R&D. Texas Instruments, Dallas, Texas

12. Wesley Pereira, "Modifing LPC Parameter Dynamics to Improve Speech Coder Efficiency", September 2001

13. Stan McClellan and Jerry D. Gibson, "Speech Signal Processing: Coding, Transmission and Storage"

14. Hwai–Tsu Hu and Hsi–Tsung Wu, "A Glottal-Excited Linear Prediction (GELP) Model for Low-Bit-Rate Speech Coding", may 1999

15. http://www.owlnet.rice.edu/~elec532/PROJECTS00/speechBIG/ Notorious LPC

16. Andreas Spanias, "Speech Coding: A Tutorial Review".

17. Andreas Spanias, "Multimedia Signal Processing Lecture Notes"

18. Schussler Mare, "Design and Simulation of a Speech Coder for Mobile Communication Systems", Master's Thesis, 1994

19. Antti Kiviluoto, "Speech Coding Standards"

20. B. S. Atal and Suzanne L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", Bell Telephone Laboratories