«ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ»

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ &
ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Αναγνώριση Προτύπων με Έμφαση στην Αναγνώριση Φωνής

Προπαρασκευή 2ης Εργαστηριακής Άσκησης

9ο Εξάμηνο

Αθανασίου Νικόλαος

03112074

Σκοπός της 2ης εργαστηριακής άσκησης είναι να δημιουργήσουμε ένα σύστημα αναγνώρισης ψηφίων από ακουστικά δεδομένα.Στο κομμάτι της προπαρασκευής παραθυριοποιούμε τα σήματα τα φιλτράρουμε σε κατάλληλο για το ανθρώπινο αυτί χώρο συχνοτήτων και υπολογίζουμε κάποια χαρακτηριστικά για κάθε σήμα φωνής που μας δίνεται(συντελεστές Cepstrum και μέσες τιμές επιμέρους και καθολικές).

Βήμα 1

Αρχικά με τη βοήθεια της εντολής *dir* η οποία μας βοηθάει να κατευθυνθούμε στο κατάλληλο φάκελο και με τις πληροφορίες που μας επιστρέφει -πιο συγκεκριμένα τα ονόματα κάθε αρχείου- τις δίνουμε ως είσοδο στην *audioread* η οποία μας επιστρέφει το σήμα της φωνής κάθε αρχείου και τη συχνότητα δειγματοληψίας.

Βήμα 2

Στη συνέχεια χρησιμοποιώνατας τη συνάρτηση *filter* και δίνοντας της τους κατάλληλους συντελεστές για αριθμητή και παρονομαστή που μας δίνονται περνάμε κάθε σήμα από το FIR φίλτρο ώστε να προκαλέσουμε μια φασματική εξομάλυνση του σήματος.

Βήμα 3

Στο παρόν βήμα παραθυριοποιούμε το σήμα με τη βοήθεια της συνάρτηση buffer χρησιμοποιώντας παράθυρο Hamming με μήκος παραθύρου 25 ms.Γενικότερα αποδεκτές τιμές είναι μεταξύ 20-40 ms ώστε από την μία να μην έχουμε μικρά παράθυρα που οδηγούν σε απώλεια μικρών αλλαγών και από την άλλη να μην έχουμε πολλές αλλαγές στις ιδιότητες του σήματος κατά τη διάρκεια του παραθύρου.Η επιλογή του παραθύρου hamming έγινε για την αποφυγή παραμόρφωσης.

Βήμα 4

Λόγω της αντίληψης του αυτιού που δεν ακολουθεί την γραμμικότητα των συχνοτήτα του fft μετασχηματισμού στο συγκεκριμένο βήμα επιχειρούμε για να εξάγουμε τα χαρακτηριστικά που μας ενδιαφέρουν από κάθε πλαίσιο να δημιουργήσουμε πρώτα ένα σύνολο φίλτρων Mel από ταοποία θα περάσουμε τα σήματα μας.Εδώ να σημειωθεί ότι στον χώρο γραμμικών συχνοτήτων χρησιμοποιήσαμε ως πάνω και κάτω όριο 300 και 8000 .Για την μετάβαση στο χώρο συχνοτήτων Mel χρησιμοποιούμε τον τύπο της εκφώνησης ώστε να

μετατρέψουμε τα δύο όρια μας και να σχηματίσουμε ένα σύνολο τιμών το f_c το οποίο στη συνέχεια χρησιμοποιώντας τη σχέση

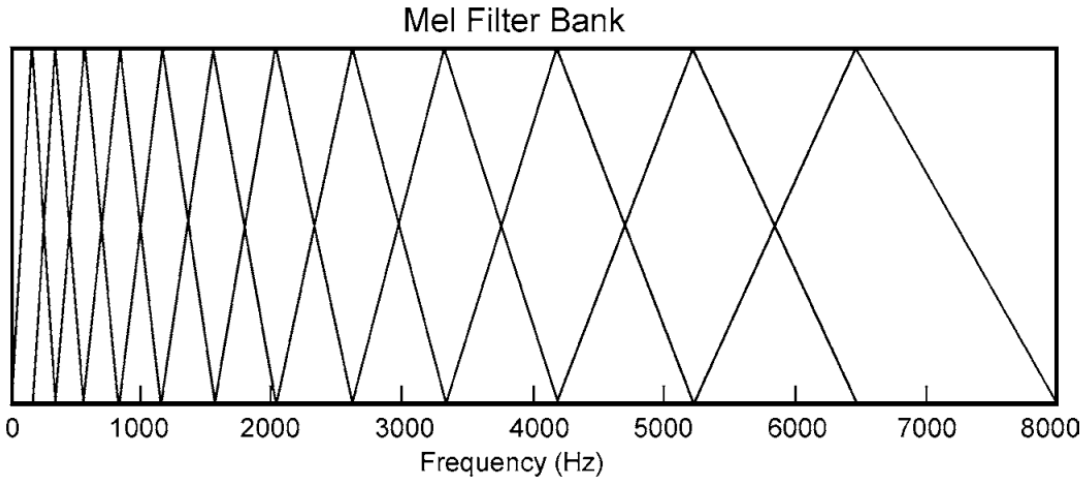
$$fmel = 700 * (10.\hat{}(fc/2595) - 1)$$

Το γυρίζουμε πίσω σε Hertz.
Και τέλος σύμφωνα με την παρακάτω σχέση

$$f = floor\left((nfft + 1) * \frac{fmel}{Fs}\right)$$

Αντιστοιχίζουμε αυτές τις παραπάνω συχνότητες στα αντίστοιχα fft bins με Fs να είναι ο ρυθμός δειγματοληψίας και nfft το μήκος του fft του σήματος μας.

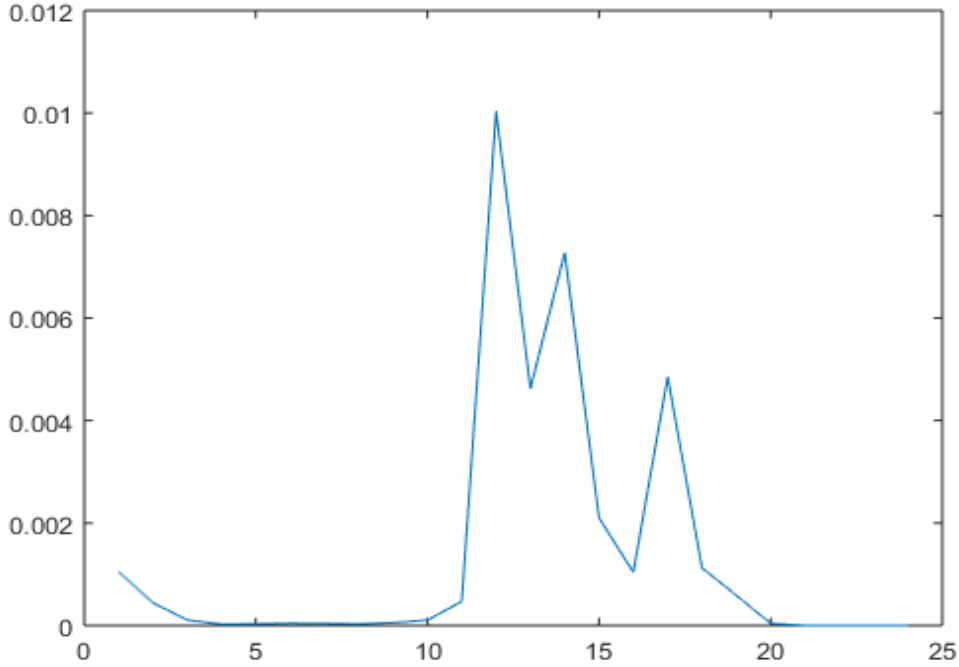
Τέλος δημιουργούμε τα φίλτρα μας τα οποία είναι τριγωνικά και πιο συγκεριμένα το πρώτο ξεκινάει στο πρώτο σημείο του συνόλου f κάνει μέγιστο(1) στο δεύτερο και μηδενίζει στο τρίτο,το δεύτερο αντίστοιχα ξεκινάει από το 2ο σημείο κάνει μέγιστο στο 3ο και μηδενίζει στο 4ο κοκ.Πιο συγκεκριμένα:



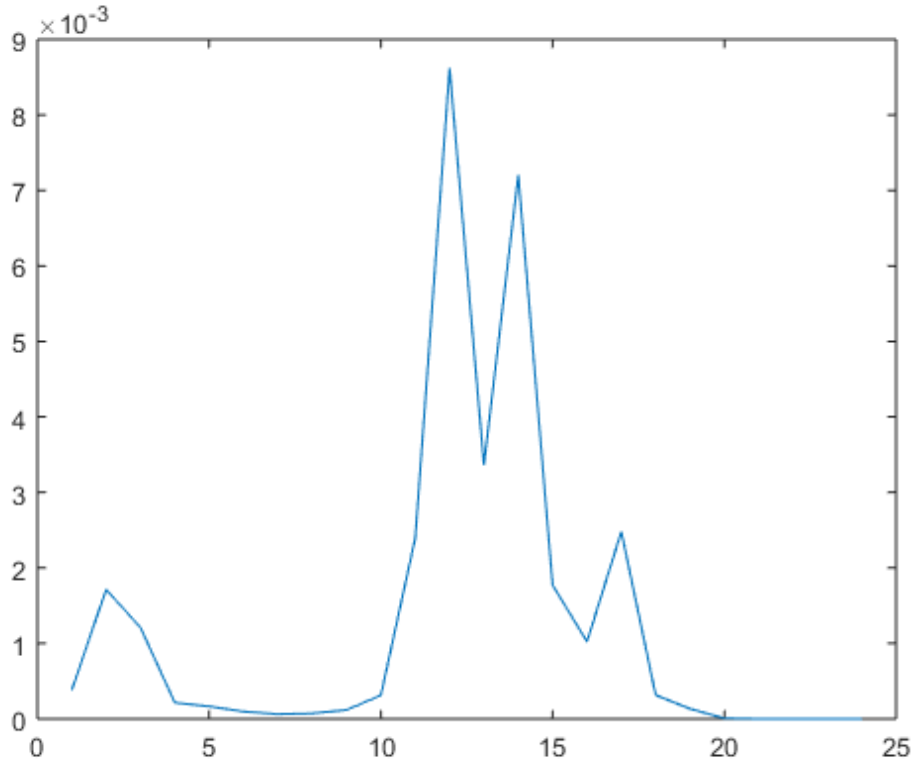
Βήμα 5

Υπολογίζουμε την ενέργεια κάθε πλαισίου μετασχήματιζοντας κατά *fft* με την ομώνυμη συνάρτηση και στη συνέχεια αθροίζοντας τα τετράγωνα των συντελεστών και διαιρώντας το άθροισμα με το μήκος του fft του κάθε παραθύρου.

Πλαίσιο 20 (eight1.wav)



Πλαίσιο 25 (eight1.wav)

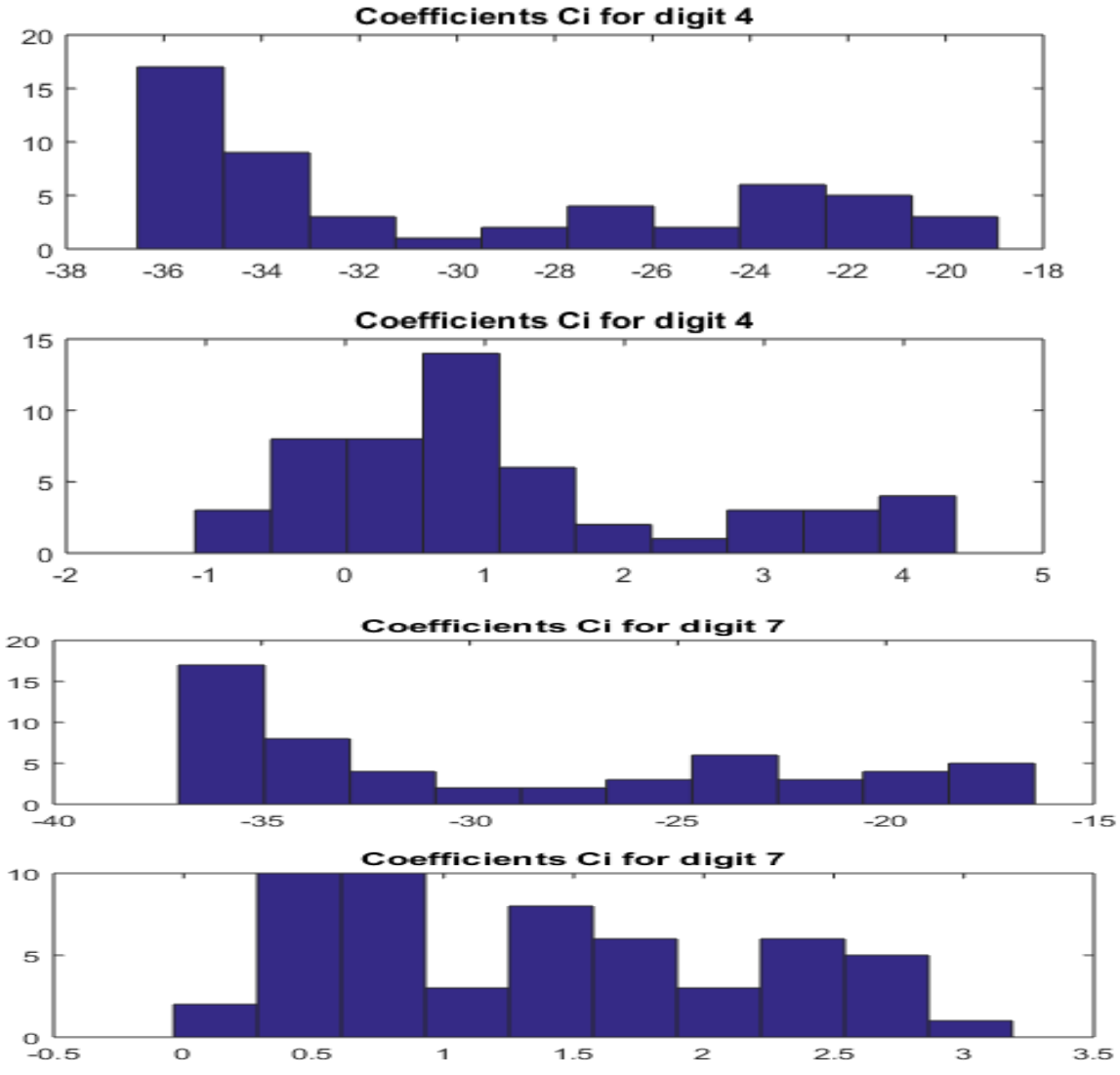


Βήμα 6

Οι συντελεστές G βρίσκονται εύκολα χρησιμοποιώντας τις ενέργειες του προηγούμενου ερωτήματος και τον τύπο που μας δίνεται στην εκφώνηση.Να σημειωθεί ότι το παρόν βήμα είναι αρκετά σημαντικό καθώς υποδηλώνει την λογαριθμική αντίληψη του ήχου από το ανθρώπινο αυτί.

Βήμα 7

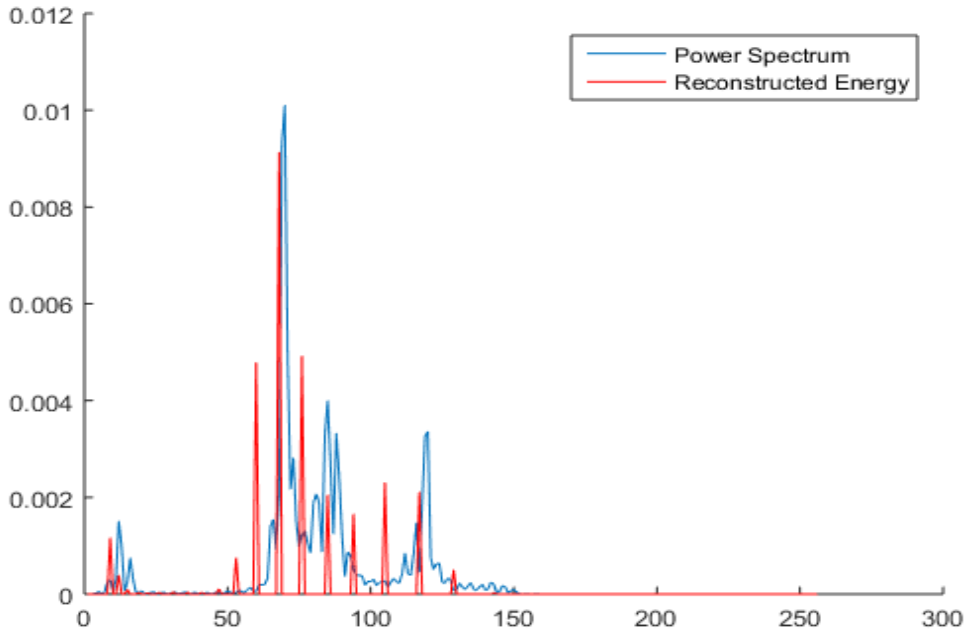
Χρησιμοποιώντας του λογαριθμικούς συντελεστές της ενέργειας του προηγούμενου ερωτήματος με τη συνάρτηση dct βρίσκουμε τους συντελεστές C και στη συνέχεια σύμφωνα με την υπόδειξη της άσκησης κρατάμε μόνο τους πρώτους 13 αφού θέλουμε 13 χρακτηριστικά για κάθε πλαίσιο.Επιλέχθηκαν σύμφωνα με τις οδηγίες της εκφώνησης τα ψηφία 7,4 προς απεικόνιση των συντελεστών τους.(03112074->2074->$\mathbf{n_1}$=27**mod**13=1 $\mathbf{n_2}$=04**mod**13=4)Πάνω ιστόγραμμα επιμέρους γραφήματος n1=1 κάτω n2=4.

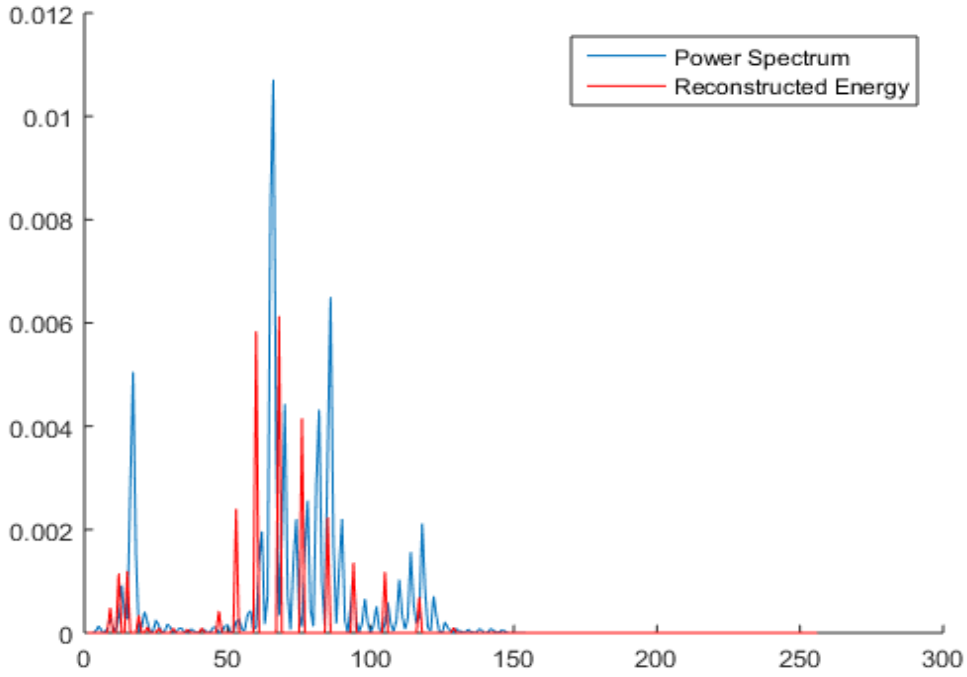


Βήμα 8

Υπολογίζουμε τον αντίστροφο dct για την εύρεση του ανακατασκευασμένου σήματος από τους συντελεστές ενέργειας και στην ίδια γραφική απεικονίζουμε το φάσμα ισχύος που προέκυψε από τον τύπο πρώτα πλαίσιο 25 μετά 20:

$$Power\ Spectrum = \frac{|fft(S(n))|^2}{no\ of\ fft\ points/2}$$

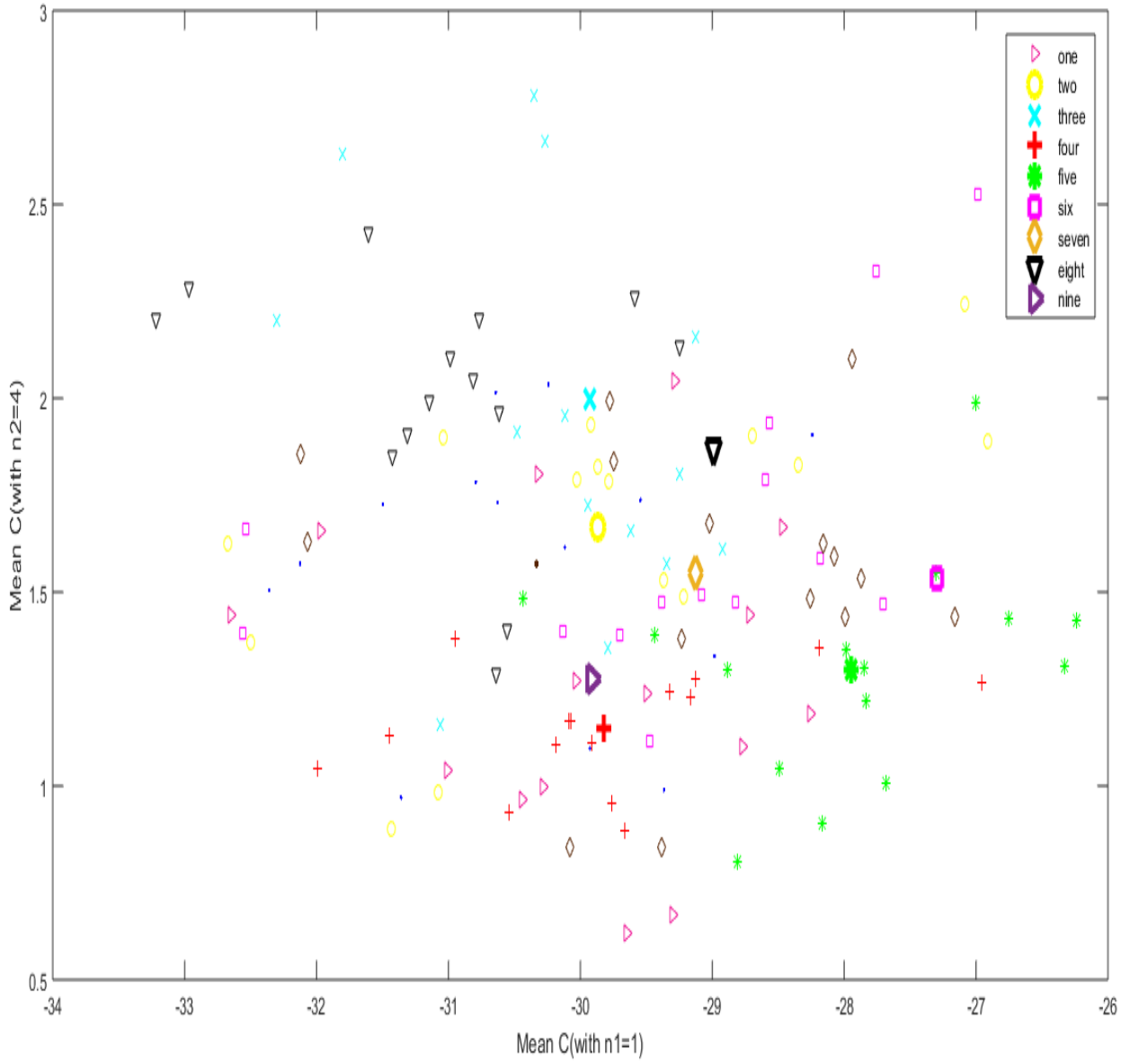




Παρατηρείται μεγάλη επιτυχία σχετικά με την ανακατασκευή των παραπάνω πλαισίων με τις όποιες ατέλειες να είναι διακιολογημένες λόγω του ότι κρατήσαμε μόνο 13 συντελεστές «Cepstrum»(C(i)).

Βήμα 9

Υπολογίζοντας τις ζητούμενες μέσες τιμές προκύπτει η παρακάτω απεικόνιση τους:



Παρατηρώντας τις μέσες τιμές οι οποίες είναι έντονα τονισμένες και μεγαλύτερες σε μέγεθος βλέπουμε ότι τα ψηφία είναι διαχωρίσιμα το ένα από το άλλο αλλά δεν μπορούμε να κατηγοριοποιήσουμε με βάση της απόστασης του καθενός από την μέση τιμή μόνο καθώς υπάρχουν και τιμές από

εκφωνήσεις που βρίσκονται πιο κοντά στη μέση τιμή κάποιου άλλου ψηφίου(πχ μια μέση τιμή του 5 ενος ομιλητή πέφτει ακριβώς πάνω στη μέση τιμή του 6 για όλες της εκφωνήσεις!).Αλλά όπως αναμέναμε παρατηρούμε ότι όλες οι μέσες τιμές για κάθε ψηφίο καταλαμβάνουν κάποιο μέρος του επιπέδου και όχι ολόκληρο.Έτσι θεωρώ ότι κάποιος αλγόριθμος clustering όπως ο kmeans θα διευκόλυνε την κατηγοριοποίηση των ψηφίων(αφού έχουμε ήδη υπολογίσει και κάποια τυπικά κέντρα(μέσες τιμές) για μια πρώτη επανάληψη).