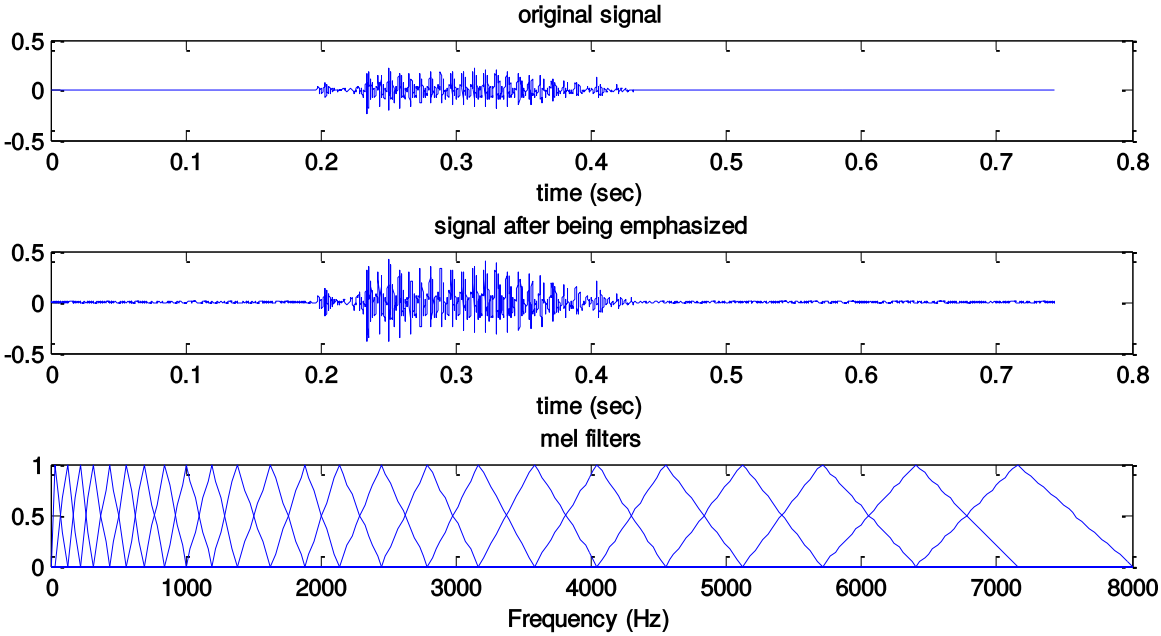
Αναγνώριση Προτύπων

Θέμα: Εξαγωγή χαρακτηριστικών απο φωνή για χρήση σε εφαρμογή αναγνώρισης

Εισαγωγή - βασικά χαρακτηριστικά

Με βάση την ανάλυση του πρώτου μέρους της εργαστηριακής άσκησης προχωρήσαμε σε τυχαία επιλογή ενός ψηφίου για υπολογισμό και εξαγωγή των βασικών χαρακτηριστικών αυτού. Συγκεκριμένα γι'αυτό το εισαγωγικό μέρος της άσκησης επιλέχθηκε το αρχείο *eight1.wav* , το οποίο αντιστοιχεί στην εκφώνηση του ψηφίου "οχτώ" από τον πρώτο ομιλητή. Έτσι πραγματοποιήθηκε η προέμφαση αυτού χρησιμοποιώντας τη συνάρτηση *filter* και η πλαισίωση αυτού σε χρονικά κομμάτια περίπου των 25ms υπολογίζοντας το πλήθος των δειγμάτων έπειτα απο απλή διαίρεση του χρόνου με την περίοδο δειγματοληψίας Ts=1/Fs (Fs=16kHz). Με αντίστοιχο τρόπο υπολογίστηκε και το χρονικό "βήμα" του κάθε παραθύρου μεταφρασμένο σε δείγματα. Στη συνέχεια κάθε χρονικό πλαίσιο πολλαπλασιάστηκε με ένα παράθυρο *hamming* του ιδίου μήκους χρησιμοποιώντας την αντίστοιχη συνάρτηση MatLab.



Σχήμα 1 : Εμφάνιση της εκφώνησης του ψηφίου "8" και των mel φίλτρων

Στη συνέχεια το ζητούμενο ήταν η κατασκευή των φίλτρων mel στον χώρο της συχνότητας γνωρίζοντας πως οι κεντρικές συχνότητες των φίλτρων αυτών στον mel χωρο συχνότητας ισαπέχουν. Δεδομένης της μαθηματικής σχέσης που συνδέει τους δύο χώρους συχνοτήτων προχωρήσαμε στην κατασκευή μιας συνάρτησης *melfilt(f,n)* η οποία θα δέχεται ως ορίσματα τις κεντρικές συχνότητες των φίτρων στο χώρο γραμμικής συχνότητας f και θα επιστρέφει των n σημείων φασμα τους. Πρώτα βέβαια θα πρέπει να υπολογιστούν οι κεντρικές συχνότητες f. Βάσει της

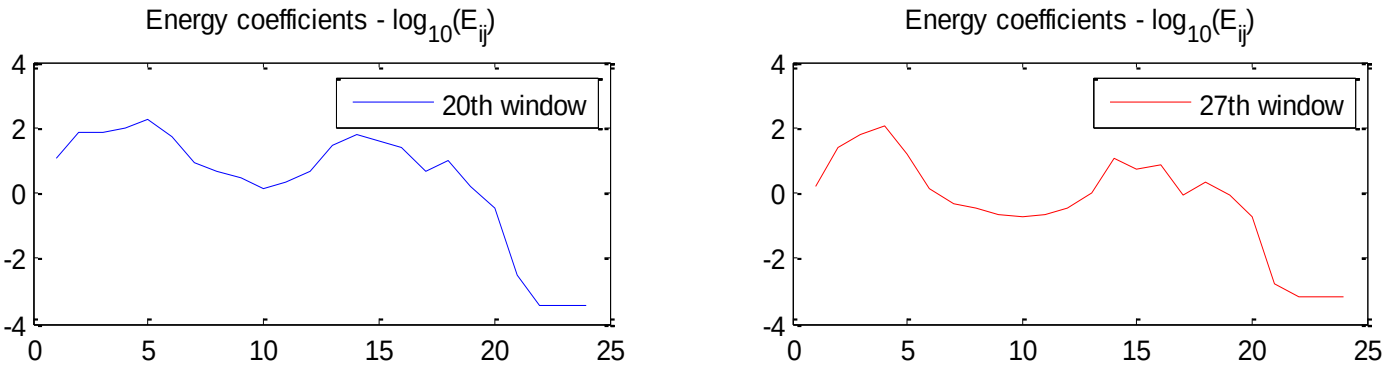
δεδομένης σχέσης και γνωρίζοντας οτι το επιθυμητό πλήθος mel φίλτρων είναι 24, υπολογίζουμε στον χώρο mel 26 το πλήθος σημεία που ισαπέχουν. Ο αριθμός 26 προκύπτει απο το γεγονός οτι θέλουμε να ξεκινήσουμε απο τη μηδενική συχνότητα χωρίς όμως να έχουμε κάποιο φίλτρο κεντραρισμένο εκεί και να φτάσουμε στη συχνότητα η οποία αντιστοιχεί στη 8kHz του γραμμικού χώρου χωρίς επίσης να δημιουργήσουμε τριγωνικό φίλτρο με αυτήν ως κεντρική συχνότητα. Έτσι τα δυο ακραία σημεία δεν θα συμπεριληφθούν υπ'οψιν. Αρκετά κατατοπιστικά είναι και τα σχόλια του κώδικα στο αντίστοιχο σημείο. Πλέον υπολογίζοντας απο τη σχέση (3) της άσκησης, λύνοντας ως προς f^j τις κεντρικές συχνότητες στον γραμμικό χώρο και χρησιμοποιώντας ως πλήθος n τα μισά σημεία για τα οποία θα εφαρμόσουμε την *fft* συνάρτηση για κάθε πλαίσιο (θα χρησιμοποιήσουμε n που αντιστοιχούν στο μισό της επόμενης δύναμης του 2 απο το πλήθος των σημείων του κάθε παραθύρου) υπολογίζουμε τα mel φίλτρα τα οποία φαίνονται στο *Σχήμα1* (η απεικόνιση εχει γίνει απο 0 εως Fs/2 ωστόσο το πλήθος των σημείων είναι 256 στη προκείμενη περίπτωση). Βασικό είναι να αναφέρουμε οτι στη συνέχεια κάθε φίλτρο επεκτάθηκε συμμετρικά στα επόμενα 256 σημεία της συχνότητας και αποθηκεύτηκε σε εναν πίνακα H μεγέθους 24x512 τέτοιο ωστε κάθε γραμμή του πίνακα να αντιστοιχεί σε ένα απο τα φίλτρα της συστοιχίας.

Στη συνέχεια κατασκευάσαμε μια συνάρτηση *energy(x,H,n,w)* η οποία δέχεται ως ορίσματα το παραθυροποιημένο πλαίσιο x, τον πίνακα mel φίλτρων H, το πλήθος των σημείων του φάσματος (512 στην προκείμενη), και το πλήθος των σημείων κάθε πλαισίου στον χρόνο w. Η έξοδος αυτής θα είναι η ενέργεια κάθε φιλτραρισμένου χρονικού παραθύρου εκμεταλεβόμενοι τη σχέση του Parseval, καθώς και ο DFT μετασχηματισμός του με απλή εφαρμογή γινομένου στη συχνότητα. Και πάλι οι ενέργειες αποθηκεύονται σε έναν πίνακα E μεγέθους px24, οπου p είναι το πλήθος των χρονικών παραθύρων, ενώ τα φάσματα των παραθύρων σε έναν πίνακα F μεγέθους px512.

Η χρήση πινάκων για την αποθήκευση της πληροφορίας κάθε παραθύρου έκανε πιο εύκολους τους υπολογισμούς καθώς επίσης αποφεύχθηκαν τα for loops σε μεγάλο βαθμό.

1.Ιδιότητες ακουστικών χαρακτηριστικών

(α') Ακολουθώντας σαφώς τις οδηγίες της άσκησης υπολογίστηκαν όπως αναφέραμε οι 24 συντελεστές ενέργειας κάθε χρονικού παραθύρου. Στην τύχη επιλέχθησαν δυό χρονικά παράθυρα των οποίων οι συντελεστές απεικονίζονται στο *Σχήμα 2* σε συνεχή γραμμή και σε λογαριθμική κλίμακα τεταγμένων. Εξηγώντας την ουσία αυτής της ενέργειας μπορούμε εύκολα να δούμε οτι κάθε συντελεστής απο 1-24 αντιπροσωπεύει την ενέργεια που είναι συσσωρευμένη εντός των συχνοτήτων του αντίστοιχου mel φίλτρου. Τα mel φίλτρα έχουν καλύτερη διακριτική ικανότητα στις χαμηλές συχνότητες ενώ καθώς η συχνότητα μεγαλώνει το εύρος ζώνης κάθε φίλτρου αυξάνεται με αποτελεσματικότητα να μειώνεται και η επιλεκτικότητα του αντίστοιχου φίλτρου.



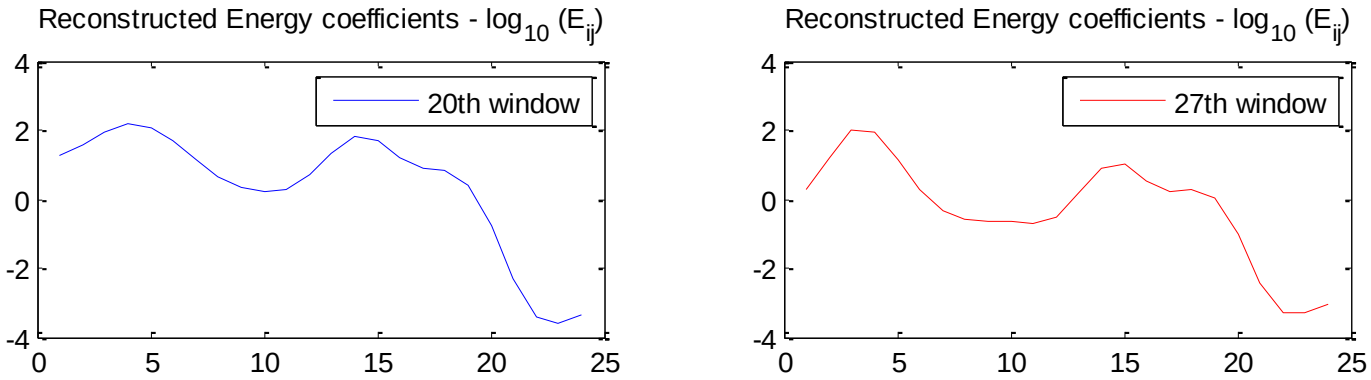
Σχήμα 2 : Απεικόνιση συντελεστών ενέργειας για δύο χρονικά παράθυρα

Επίσης απο τη σχέση του Parseval για πεπερασμένα σήματα διακριτού χρόνου έχουμε οτι:

$$\sum_{n=0}^{N-1}|x(n)|^2 = \frac{1}{N}\sum_{k=0}^{N-1}|X(k)|^2$$

Βάσει της παραπάνω σχέσης ο υπολογισμός έγινε απ'ευθείας στον χώρο της συχνότητας χωρίς ωστόσο να χρειάζεται κάποιος παράγοντας κανονικοποίησης αφού στην εν λόγω άσκηση ενδιαφέρει περισσότερο η σχετική σύγκριση ενέργειας απο παράθυρο σε παράθυρο. Περισσότερες λεπτομέρειες υπολογισμού είναι εμφανής εντός του παραδωτέου κώδικα.

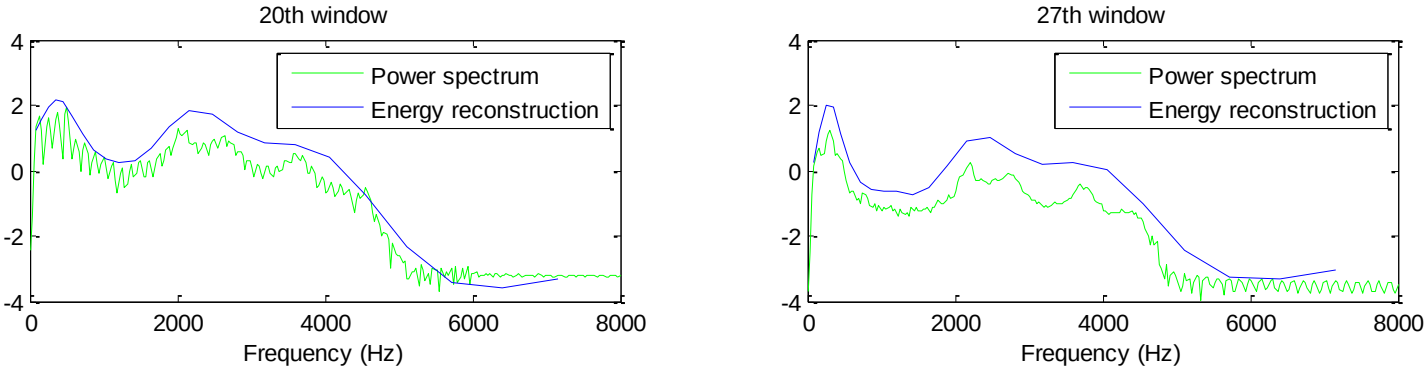
(β') Αφού πρώτα υπολογίσουμε τους DCT συντελεστές και κρατήσουμε του πρώτους 13 όπως αναφέρουν τα βήματα της άσκησης, κατασκευάσαμε μια συνάρτηση *dct_reconstruct(C,n,w)* η οποία δέχεται σαν ορίσματα τον πίνακα των DCT συντελεστών C (σε κάθε γραμμή έχει τους DCT συντελεστές του αντίστοιχου παραθύρου), το πλήθος των συντελεστών ενέργειας για ανακατασκευή n (24 στην περίπτωσή μας), και το πλήθος των χρονικών πλαισίων της εκφώνησης w, ενω στην έξοδο επανεκτιμά τους 24 συντελεστές ενέργειας για κάθε πλαίσιο. Οι αντίστοιχες γραφικές παραστάσεις φαόινονται στο *Σχήμα 3.*



Σχήμα 3 : Απεικόνιση ανακατασκευασμένων συντελεστών ενέργειας για δύο χρονικά παράθυρα

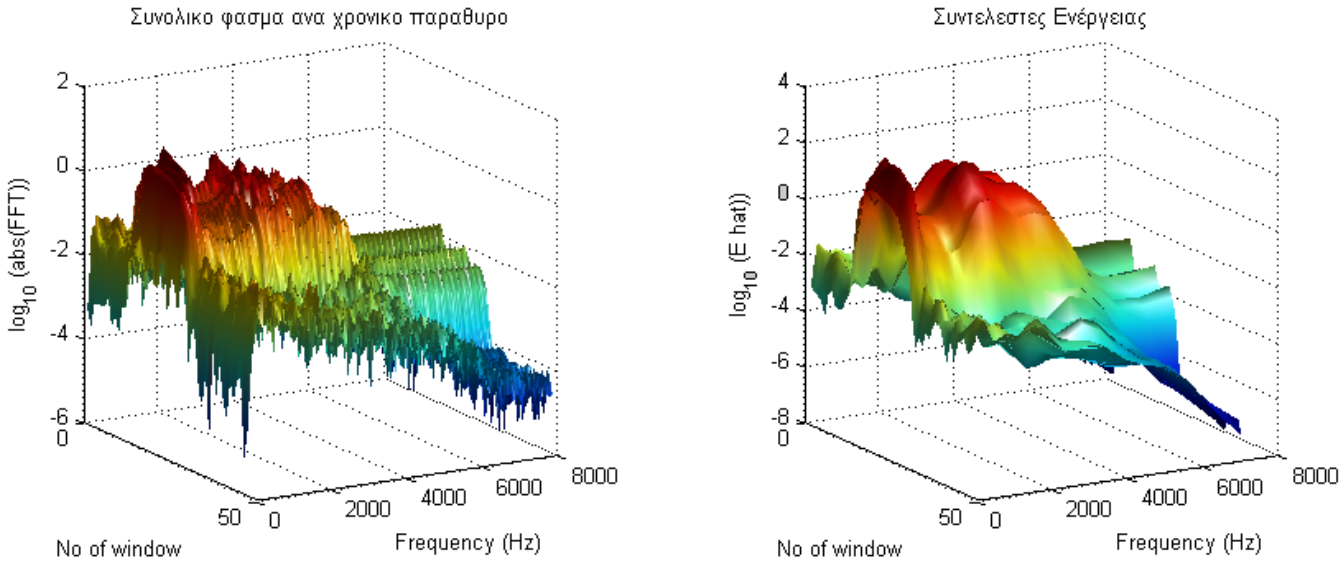
Άν συγκρίνουμε τα σχήματα 3 και 4 μπορούμε να διαπιστώσουμε οτι πλέον στους ανακατασκευασμένους συντελεστές οι μορφές των καμπύλων είναι πιο ομαλές. Αυτό με τη σειρά του σημαίνει οτι οι μεταβολές των συντελεστών ενέργειας είναι πιο ομαλές και αυτό είναι λογικό γιατι αν φανταστουμε τους 24 συντελεστές του *Σχήματος 2* για κάθε περίπτωση σαν ένα διαφορετικό σήμα, ουσιαστικά εφαρμόσαμε ενα βαθυπερατό φιλτράρισμα στον χώρο της συχνότητας μηδενίζοντας τις συνιστώσες απο την 14η και πάνω μεσω του πίνακα DCT.

(γ') Έχοντας υπολογίσει το φάμα ισχύος κάθε παραθυρωμένου πλαισίου μπορούμε πλέον να απεικονίσουμε τις παραπάνω γραφικές παραστάσεις με τα φάσματα αυτά ταυτόχρονα. Αυτό που θέλει προσοχή σε αυτήν την περίπτωση είναι η απεικόνιση των παραπάνω γραφικών που προκύπτουν απο τους συντελεστές ενέργειας με τον σωστό τρόπο στο χώρο γραμμικής συχνότητας. Κατι τέτοιο όμως δεν είναι δύσκολο καθώς έχουμε ήδη υπολογίσει τα 24 κέντρα των συχνοτήτων των mel φίλτρων και το μόνο που χρειάζεται είναι η αντιστοιχία των παραπάνω συντελεστών στα κέντρα αυτά. Επίσης για την εφάνιση χρειάζεται να προσθέσουμε και κάποια DC συνιστώσα για να μπορέσει να απεικονιστεί η ενέργεια ως περιβάλλουσα του φάσματος. Τα αποτελέσματα φαίνονται στο *Σχήμα 4.*



Σχήμα 4 : Απεικόνιση ενέργειας με φάσμα ισχύος μαζί

Με αυτόν τον τρόπο μπορούμε να δούμε οτι η ενέργεια ακολουθεί την γενικότερη μορφή του φάσματος χωρίς να έχει την πολυπλοκότητα του και χωρίς να λαμβάνει υπόψην τις πολύ απότομες μεταβολές, χαρακτηριστικά που οφείλονται κυρίως στον ομιλητή (pitch) και όχι στην πληροφορία που θέλουμε να λάβουμε υπ'οψιν. Στη συνέχεια στο *Σχήμα 5* απεικονίζεται το φάσμα ισχύος αλλα και η ανακατασκευασμένη ενέργεια, για κάθε χρονικό παράθυρο.

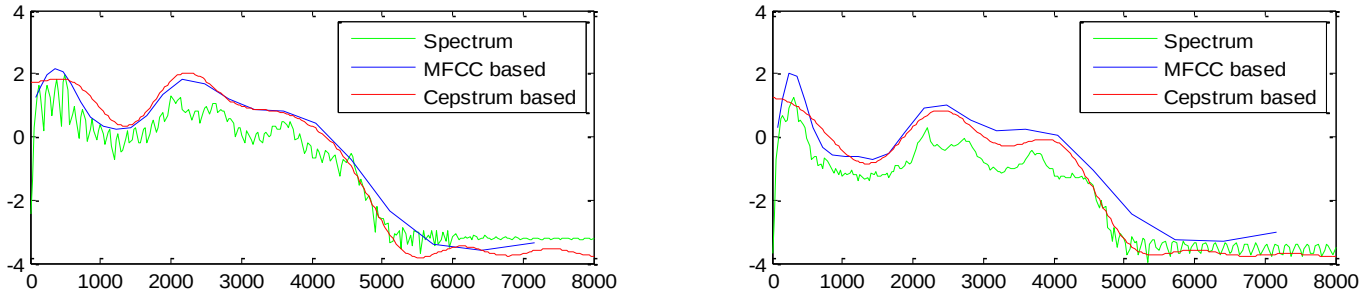


Σχήμα 5 : Απεικόνιση ενέργειας με φάσμα ισχύος μαζί για καθε παραθυρωμενο πλαίσιο

Απο τα παραπάνω ενδιαφέροντα γραφήματα επιβεβαιώνουμε οτι η ιδιότητα της υπολογιζόμενης ενέργειας ως φασματική περιβάλλουσα ισχύει για κάθε πλαίσιο καθώς μπορούμε να πούμε οτι η δεξιά επιφάνεια του σχήματος μπορεί να "σκεπάσει" τη φασματική επιφάνεια των χρονικών παραθύρων το οποίο είναι ουσιαστικά ενα διαφορετικού είδους φασματογράφημα εφ'οσον τα χρονικά παράθυρα προς μελέτη είναι επικαλυπτόμενα.

(δ') Στη συνέχεια υπολογίζουμε τους συντελεστές του cepstrum απο τις γνωστές σχέσεις ή και έτοιμες συναρτήσεις του matlab, άν και στην εν λόγω περίπτωση έγινε manually γιατί θέλαμε το ίδιο πλήθος σε σημεία με τον fft για θέματα συμβατότητας κατα την απεικόνιση. Ως γνωστόν κρατώντας μόνο τις πρώτες συνιστώσες ουσιαστικά λαμβάνουμε υπ'όψιν μονο την πληροφορία του ηχητικού σωλήνα χωρίς να μας ενδιαφέρουν οι υψηλές συχνότητες που ωφείλονται στο pitch του ομιλούντα. Γυρίζοντας απο τον χωρο του cepstrum ξανα στο πεδίο της συχνότητας θα έχουμε την φασματική

περιβάλλουσα που είναι παρόμοια με αυτή που προκύπτει απο τους MFCCs όπως φαίνεται και στο *Σχήμα 6.*

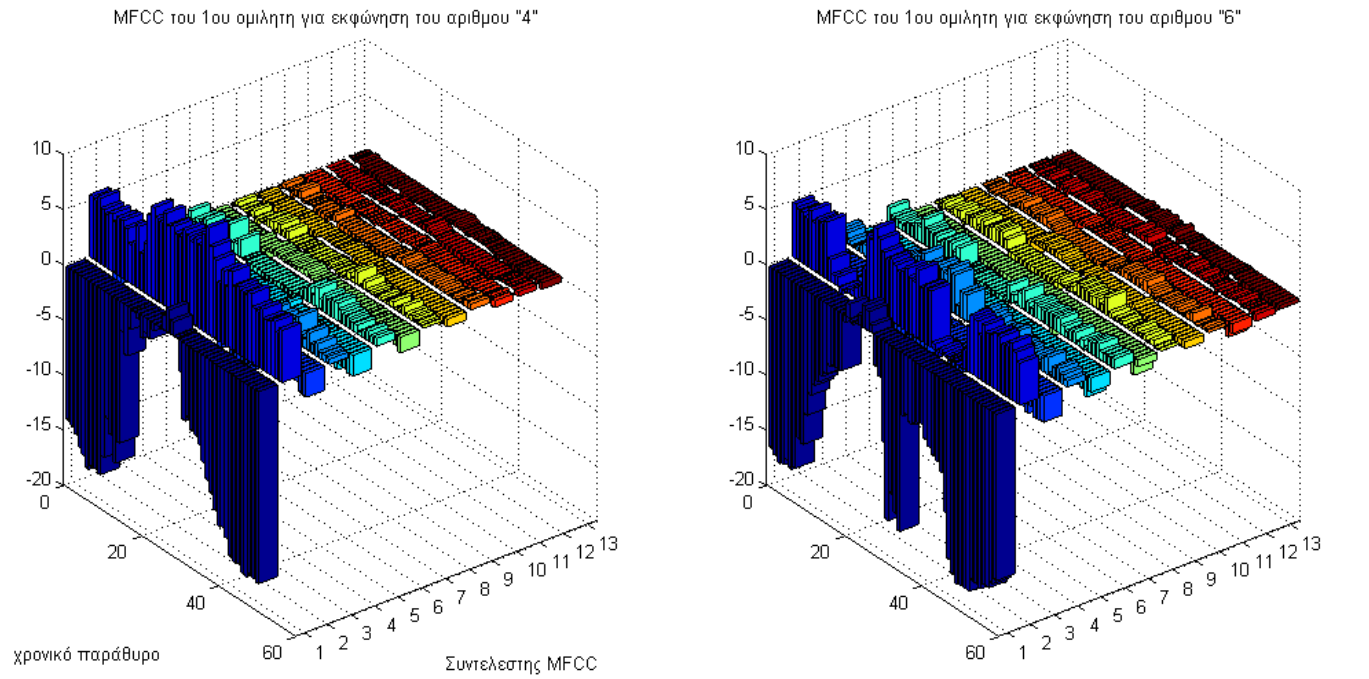


Σχήμα 6 : Απεικόνιση MFCC ενέργειας, φάσματος ισχύος και φασματος απο cepstrum

Όπως αναμένουμε οι διαφορές είναι σχετικά μικρές και βρίσκονται κυρίως στις μικρότερες και στις μεγαλύτερες τιμές συχνοτήτων. Αυτό δικαιολογείται με το γεγονός οτι τα mel φίλτρα προκύπτουν απο διάφορες ψυχομετρικές μετρήσεις που έχουν γίνει σε ανθρώπους και όπως φαίνεται η ανθρώπινη αντίληψη στις υψηλότερες και χαμηλότερες συχνότητες δεν συμβαδίζει άριστα με τη μαθηματική ερμηνεία των χαρακτηριστικών του ήχου, καθώς ο όρος της υποκειμενικότητας είναι προφανώς αισθητός. Όλα τα παραπάνω που περιφράφτηκαν στο 1ο μέρος υλοποιούνται μέσω του παραδωτέου script ***meros1.m***

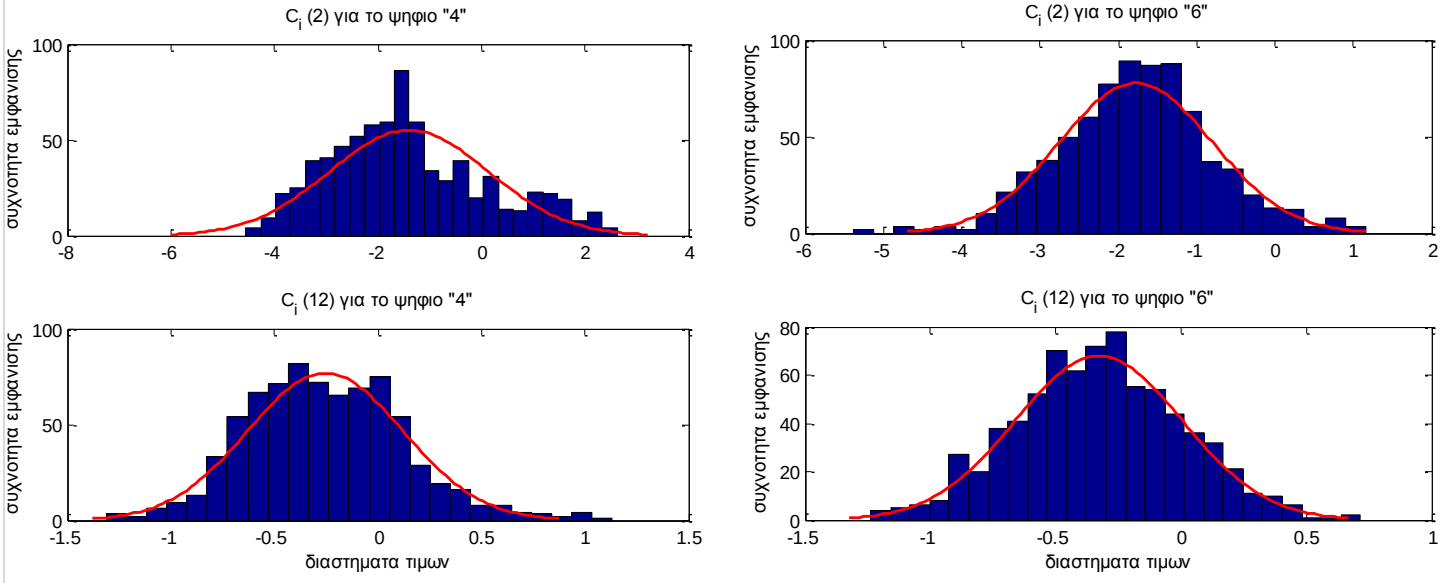
2.Στατιστικά χαρακτηριστικά ακουστικών συχνοτήτων

Σ'αυτό το στάδιο κατασκευάστηκαν δυο βοηθητικές συναρτήσεις για τον ευκολότερο υπολογισμό των ζητούμενων. Η πρώτη, η *C=MFCC(s)* , δέχεται ως όρισμα ένα σήμα εισόδου, π.χ *s=eight1.wav* , και στην έξοδο επιστρέφει τον πίνακα C ο οποίος έχει σε πλήθος τόσσες γραμμές όσες και το πλήθος των παραθυρωμένων πλαισίων του σήματος εισόδου, ενω το πλήθος των στηλών είναι δεκατρείς, όσοι και οι mel frequency cepstrum coefficients που θέλουμε να κρατήσουμε. Έτσι στην k-γραμμή του πίνακα εξόδου υπάρχουν οι 13 συντελεστές C_i του k-οστού παραθυρωμένου πλαισίου της εκφώνησης *s*. Στο συγκεκριμένο μέρος τα ψηφία που αναλύθηκαν βάσει του αριθμού μητρώου είναι τα k_1=4 και k_2=6. Η δεύτερη συνάρτηση είναι η *[C_all , index]=read _all_speakers(d)* η οποία δέχεται ως ορισμα έναν ακαίραιο *d* ο οποίος αντιστοιχεί στο ψηφίο που θέλουμε να διαβάσουμε (π.χ d=4) και επιστρέφει με χρήση της *MFCC* που μόλις αναφέραμε τους 13 πρώτους mel frequency cepstrum coefficients για κάθε παραθυρωμένο πλαίσιο κάθε εκφώνησης στον πίνακα *C_all*. Στις πρώτες n-γραμμές περιέχονται οι συντελεστές της πρώτης εκφώνησης, στις επόμενες m-γραμμές οι συντελεστές της δεύτερης εκφώνησης και ούτω καθεξής. Ο *index* είναι ενα array το οποίο περιέχει τις θέσεις στις γραμμές του C_all στις οποίες έχουμε αλλαγή εκφώνησης ωστε να μπορέσουμε στη συνέχεια να μελετήσουμε κάθε εκφώνηση ξεχωριστά. Βάσει όλων αυτών μπορούμε αρχικά να παρατηρήσουμε στο *Σχήμα 7* τους συντελεστές C_i για έναν μονο εκφωνητή ξεχωριστά με χρήση της συνάρτησης *bar3.* Συγκεκριμένα στο σχήμα αυτό απεικονίζονται οι συντελεστές για τις εκφωνήσεις των ψηφίων "τέσσερα" και "έξι" απο τους πρώτους εκφωνητές.



Σχήμα 7 : Συντελεστές MFCC για καθε παραθυρωμένο πλαίσιο της πρώτης εκφώνησης των ψηφίων "4" και "6"

Απο τα δυο παραπάνω διαγράμματα παρατηρούμε οτι ο πρώτος και ο δεύτερος mfcc (δηλαδή για n=0 και n=1) έχουν μεγαλύτερες (κατ'απόλυτον) τιμές απο τους υπόλοιπους. Αυτό μας κάνει να αναμένουμε πως θα είναι και πιο συμαντικοί κάτι όμως που θα επαληθευτεί στα επόμενα ερωτήματα. Χαρακτηριστικό είναι πως ο πρώτος συντελεστής παίρνει αρκετά αρνητικές τιμές στα άκρα των εκφωνήσεων, εκέι δηλαδή που δεν υπάρχει πολύ σημαντική πληροφορία όπως φαίνεται και απο το *Σχήμα 1*, ενώ εκέι που η ενέργεια είναι υψηλότερη οι τιμές γίνονται λιγότερο αρνητικές, έως και θετικές.Στη συνέχεια απο τον πίνακα *C_all* όπως αυτός περιγράφτηκε παραπάνω, μπορούμε να δούμε τα ιστογράμματα των συντελεστών C(2) και C(12) για το σύνολο των πλαισίων και των εκφωνήσεων καθώς αυτοί ουσιαστικά βρίσκονται στην 3η και στην 13η στήλη του πίνακα C_all αντίστοιχα. Στο *Σχήμα 8* φαίνονται τα ζητούμενα ιστογράμματα με χρήση της συνάρτησης *histfit* για 25 διαστήματα.

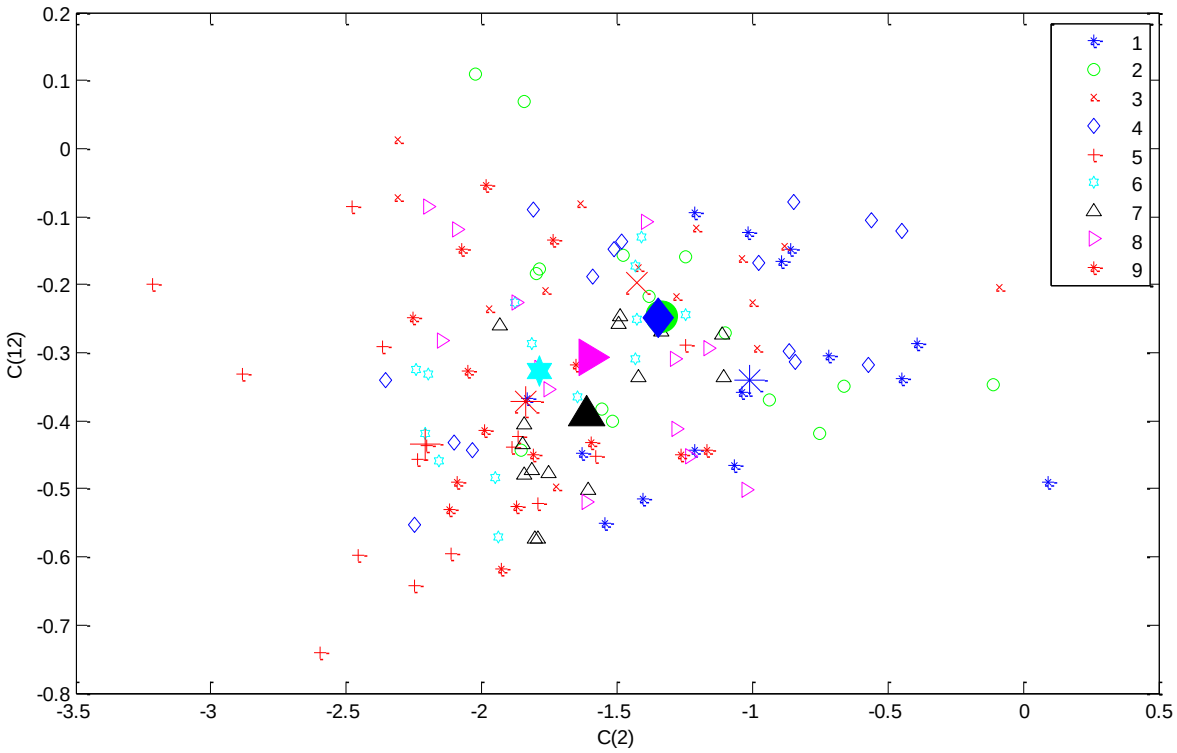


Σχήμα 8 : Ζητούμενα ιστογράμματα για το ψηφίο "4" και το ψηφίο "6"

Η συνάρτηση *histfit* που χρησιμοποιήθηκε κάνει ταυτόχρονα μια προσέγγιση των δεδομένων με κανονική κατανομή. Απο τα παραπάνω ιστογράμματα βλέπουμε ωστόσο οτι οι μέσες τιμές των συντελεστών C(2) είναι σχετικά ίσες για τα δύο ψηφία, όπως και των συντελεστών C(12). Κάτι τέτοιο μας οδηγεί στο σκεπτικό οτι ο 3ος και ο 13ος συνελεστής mfcc δεν θα μπορούσαν να αποτελέσουν "καλούς" συντελεστές για διαχωρισμό των ψηφίων, κάτι που συμβαδίζει με όσα θα δούμε στη συνέχεια. Όλο το 2ο μέρος που περιγράφθηκε παραπάνω υλοποιείται μέσω του script **meros2.m** που βρίσκεται εντός του παραδωταίου κώδικα με κατατοπιστικά σχόλια.

3.Στατιστικά χαρακτηριστικά των C(2) και C(12) για όλες τις εκφωνήσεις

Στη συνέχεια σ'αυτό το μέρος προχωρήσαμε ένα βήμα παραπάνω απ'οτι προηγουμένως και υπολογίσαμε τους μέσους όρους των συντελεστών C(2) και C(12) για κάθε εκφώνηση και για κάθε ψηφίο. Αυτο επιτεύχθηκε με τη δημιουργία της συνάρτησης *AverageC(n)* η οποία επιστρέφει 9 διαφορετικούς πίνακες Mi_n κάθε ένας απο τους οποίους περιέχει τους μέσους όρους των C(n) συντελεστών του i-ψηφίου για κάθε εκφώνηση ξεχωριστά. Έτσι απεικονίζοντας τα ζεύγη (C(2),C(12)) που προκύπτουν απο κάθε εκφώνηση στο επίπεδο προκύπτει το *Σχήμα 9*.

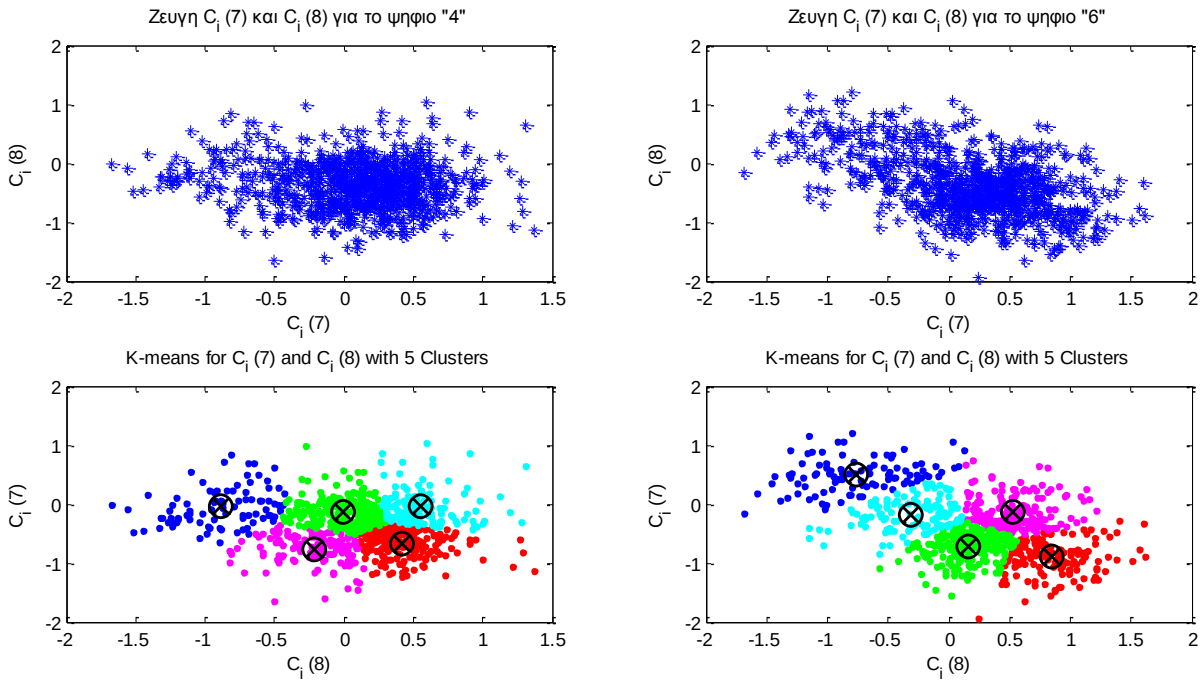


Σχήμα 9 : Συντελεστές C(2) και C(12) για κάθε εκφώνηση και κάθε ψηφίο ξεχωριστά.

Στο παραπάνω σχήμα με μεγαλύτερο μέγεθος φαίνονται και οι συνολικοί μέσοι όροι των C(2) και C(12) για κάθε ψηφίο. Τα σύμβολα για κάθε ψηφίο φαίνονται στο πάνω δεξιά μέρος του σχήματος. Παράτηρούμε λοιπόν οτι ο διαχωρισμός των ψηφίων με αυτή τη μέθοδο θα ήταν εντελώς αναξιόπιστος όπως ήταν αναμενόμενο. Ίσως μόνο για δύο ψηφία, το "5" και το "2" των οποίων οι μέσοι όροι απέχουν περισσότερο απο των άλλων ζευγών θα μπορούσε να γίνει διαχωρισμός αλλα σε καμία περίπτωση δεν θα μπορούσαμε να εξάγουμε ένα καθολικό μοντέλο. Όλα τα παραπάνω υλοποιούνται μέσω του script **meros3.m**

4.Ομαδοποίηση χαρακτηριστικών

Και πάλι με τη χρήση της *read_all_speakers(d)* είναι πλέον εύκολο να κρατήσουμε την 8η στήλη που αντιστοιχεί στο συντελεστή C(7) και την 9η στήλη για τον συντελεστή C(8). Στη συνέχεια απεικονίζουμε και τους δυο στο επίπεδο και με χρήση της συνάρτησης *kmeans* ομαδοποιήσαμε τα ζεύγη αυτά σε 5 κλάσεις. Οι παράμετροι που επιλέχθησαν στη kmeans ήταν η ευκλείδια απόσταση, batch μέθοδος, 10 ρέπλικες και ως αρχικοποίηση τυχαία ζεύγη απο τα δείγματα. Τα αποτελέσματα είναι εμφανή στο *Σχήμα 10.*

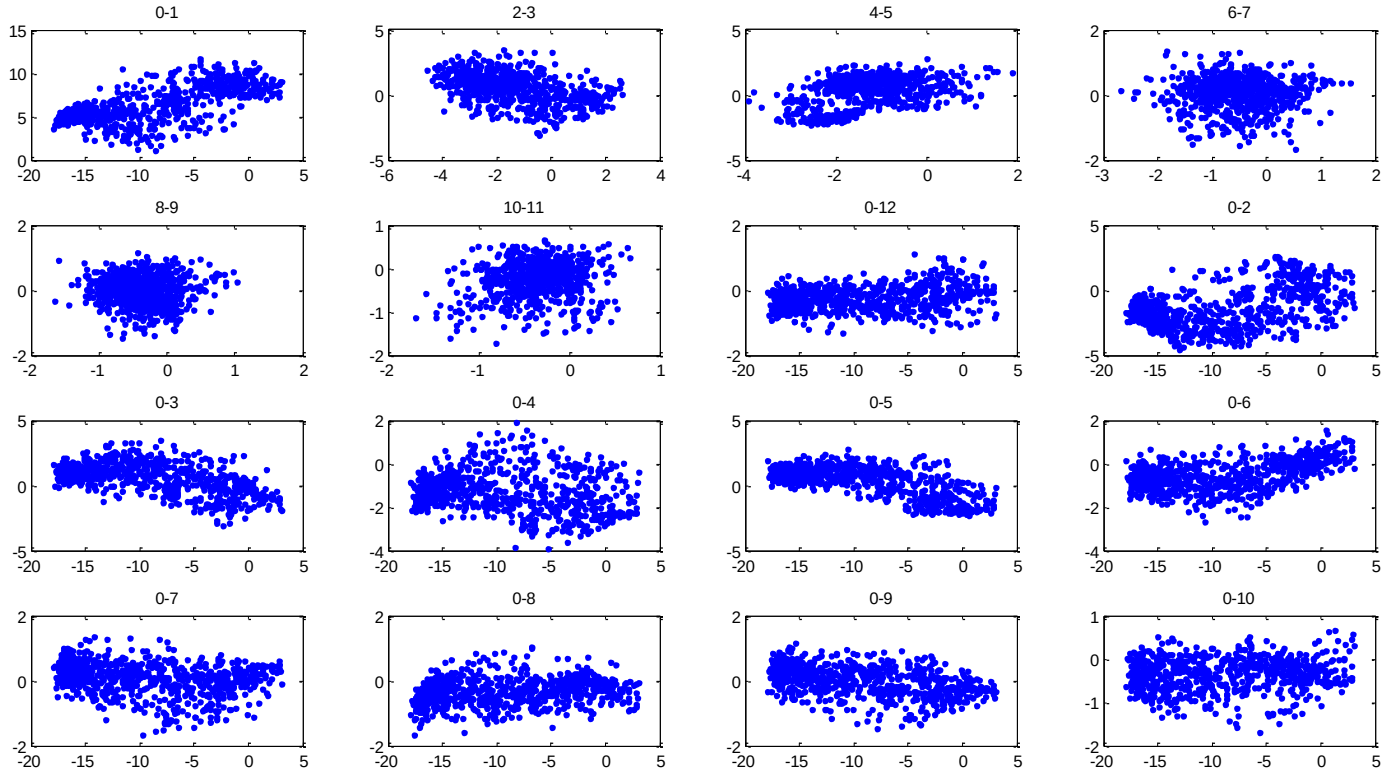


Σχήμα 10 : clustering με kmeans για τους συντελεστές C(7) και C(8)

Απο τα παραπάνω διαγράμματα παρατηρούμε οτι τα ζεύγη των συντελεστών είναι αρκετά κοντά το ένα με το άλλο και εκ'πρώτης όψεως απο την πρώτη σειρά του σχήματος δεν φαίνεται οτι μπορούν να ομαδοποιηθούν. Φυσικά ο kmeans αλγόριθμος θα δώσει αποτέλεσμα το οποίο φαίνεται στη δεύτερη σειρά του σχήματος χωρίς ωστόσο να σημαίνει οτι αυτή η ομαδοποίηση έχει κάποια αξιοπιστία όσον αφορά το συνολικό διαχωρισμό των κλάσεων των πλαισίων. Αναμένουμε οτι οι C(7) και C(8) δεν είναι ικανοί να δημιουργήσουν ενα αξιόπιστο μοντέλο καθώς οι κλάσεις είναι πολύ κοντά μεταξύ τους με εξαίρεση την κλάση που αντιστοιχεί σε μπλέ χρώμα. Όλα τα παραπάνω υλοποιούνται απο το script **meros4.m**

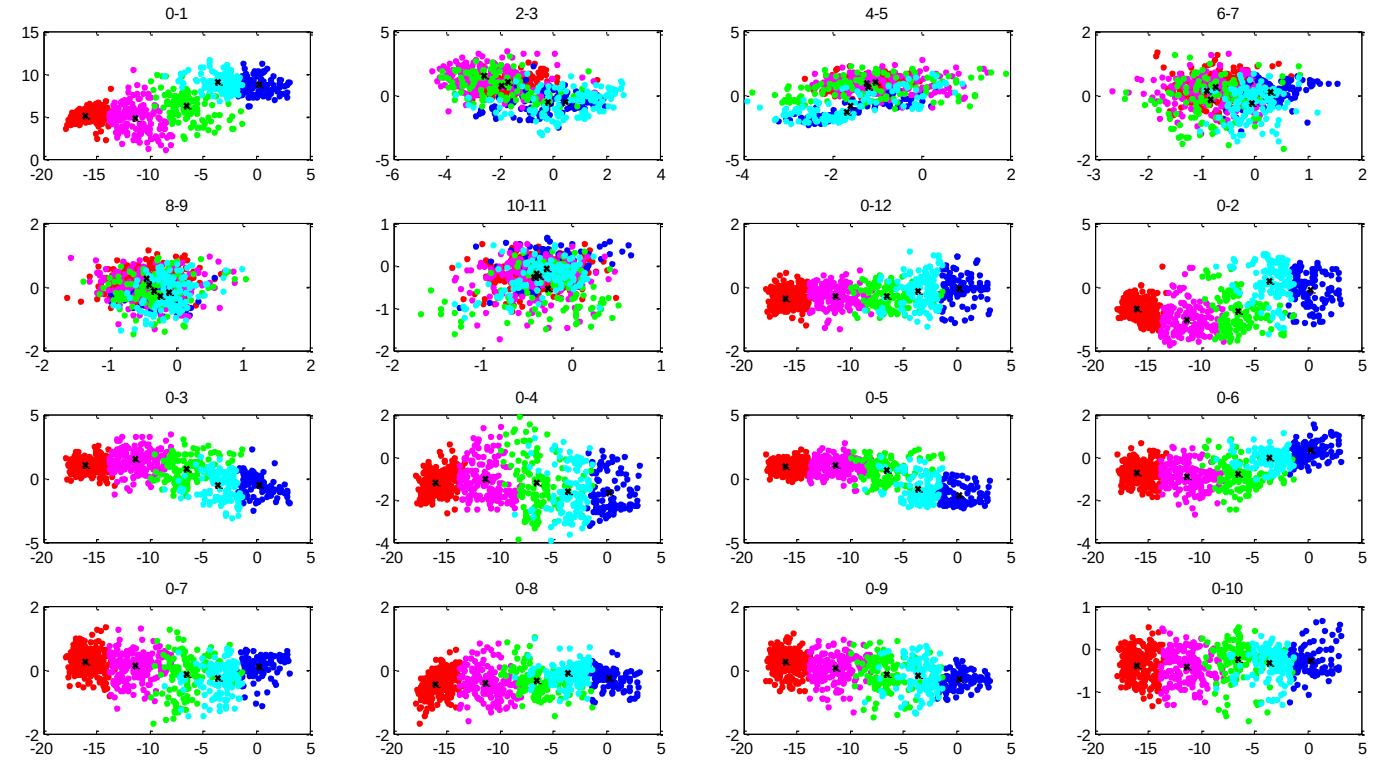
5.Ομαδοποίηση χαρακτηριστικών και χρονική εξάρτηση

Στη συνέχεια απο τον πίνακα C_all κρατάμε όλες τις στήλες ωστε να λάβουμε όλους τους συντελεστές υπ'όψιν στην κατηγοριοποίηση οπότε κάθε παράθυρο χαρακτηρίζεται απο ένα 13D διάνυσμα. Εφαρμόζουμε και πάλι τον αλγόριθμο kmeans στον 13D χώρο και απεικονίζουμε τα αποτελέσματα σε επιλεγμένες απεικονίσεις σε 2D υποχώρους του για παρατήρηση.



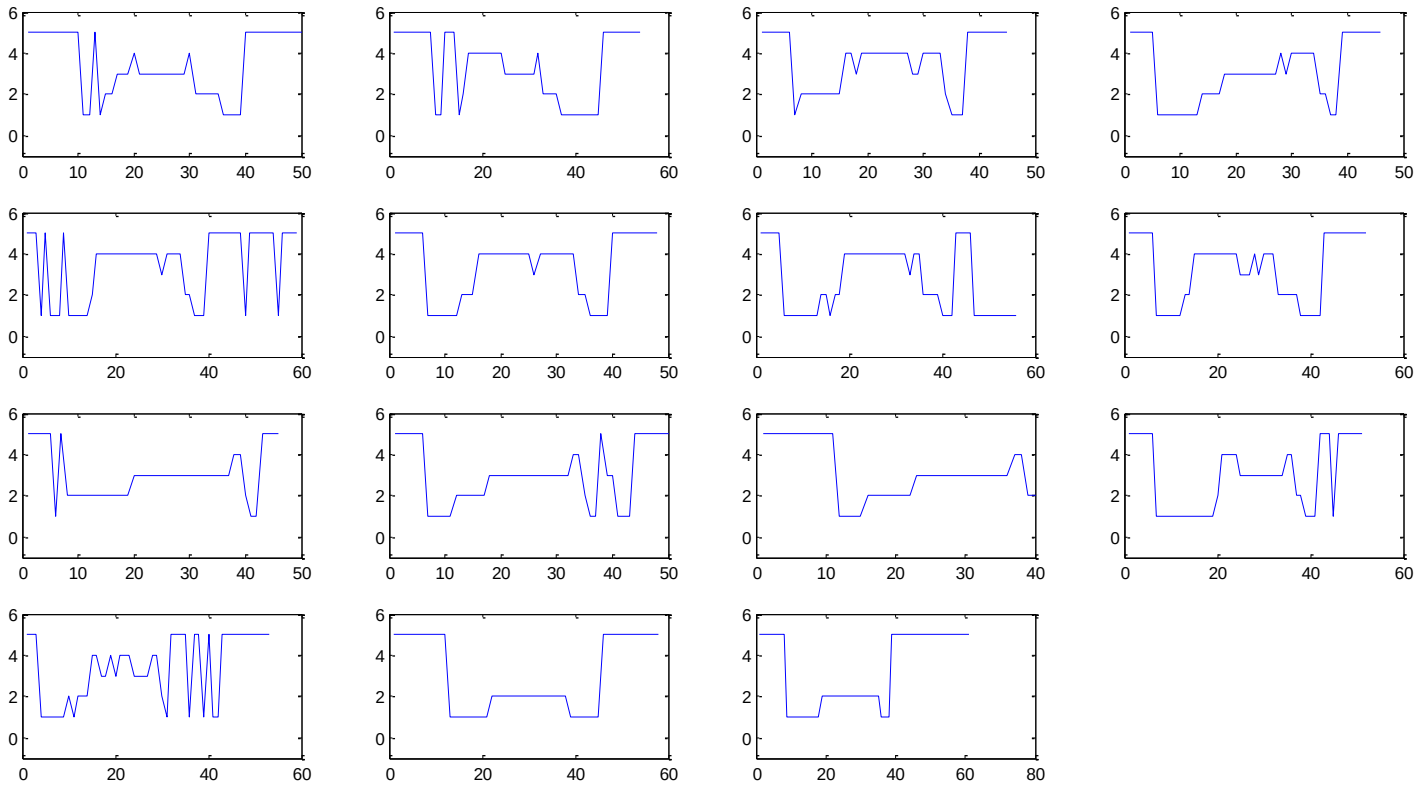
Σχήμα 11: απεικόνιση ζευγών mfcc για το ψηφίο "4"

Στο παραπάνω σχήμα φαίνεται η προβολή των 13D MFCCs σε 2D επίπεδα. Σε κάθε απεικόνιση αναφέρεται ο δείκτης *n* , π.χ στο πρώτο διάγραμμα το "0-1" σημαίνει οτι απεικονίζονται τα ζεύγη που αντιστοιχούν στον 1ο (n=0) και στον 2ο (n=1) συντελεστή mfcc. Στο *Σχήμα 12* εμφανίζονται και πάλι οι απεικονίσεις αυτές μετά την ομαδοποίηση του 13D χώρου με χρήση του kmeans αλγόριθμου για 5 clusters.



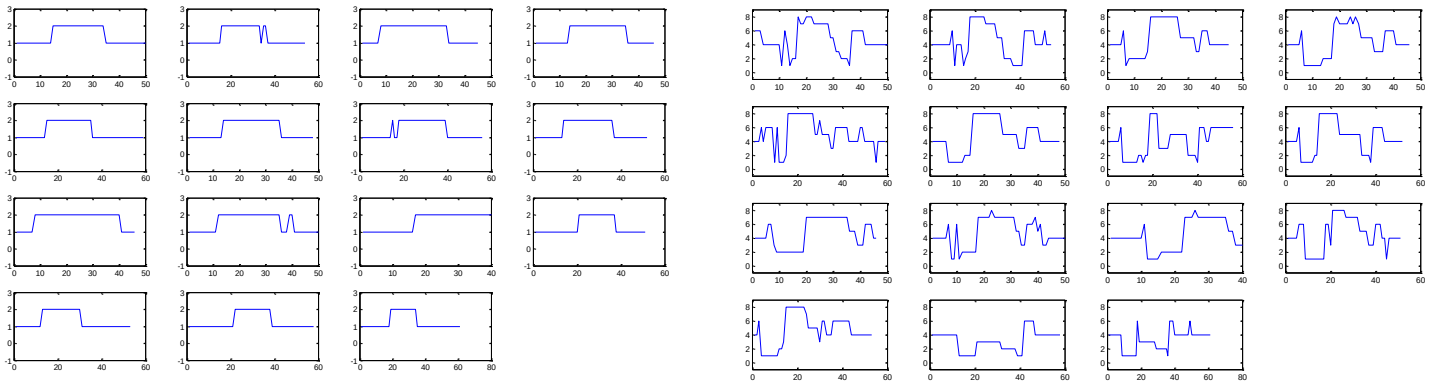
Σχήμα 12: απεικόνιση των clusters μετά απο ομαδοποίηση με kmeans (ψηφίο "4")

Το βασικότερο συμπέρασμα που προκύπτει απο τα παραπάνω διαγράμματα είναι η σημαντικότητα του πρώτου συντελεστή mfcc. Κάτι τέτοιο είναι εμφανές καθώς στα ζεύγη μεταξύ του πρώτου συντελεστή (για n=0) και κάποιου άλλου απο τους υπόλοιπους, οι ομάδες μπορούν να παρατηρηθούν με ευκολία, ενώ σε άλλη περίπτωση παρατηρώντας δυο άλλα ζεύγη (π.χ 2-3 μεταξύ 3ου και 4ου συντελεστή) η προβολή των κλάσεων στο επίπεδό τους δεν επιτρέπει τον οπτικό διαχωριμό των κλασεων που προκύπτουν απο τον kmeans αλγόριθμο. Αυτό συμβαίνει καθώς ο πρώτος συντελεστής είναι και ο πιο σημαντικός κατα τη διαδικασία ομαδοποίησης κάτι εξάλλου που ήταν αναμενόμενο με βάση όσα παρατηρήσαμε στο 2ο μέρος και το *Σχήμα 7*. Στη συνέχεια για κάθε εκφώνηση του ψηφίου "4" μπορούμε να δούμε σε ποιά κλάση ανήκει κάθε πλαίσιο αυτού παρατηρώντας ουσιαστικά την κίνηση των πλαισίων στον χώρο των τάξεων. Μετά απο αρκετές δοκιμές προέκυψε οτι ο βέλτιστος αριθμός κλάσεων για παρατήρηση μιας κοινής μορφής κατηγοριοποίησης των πλαισίων είναι 5.



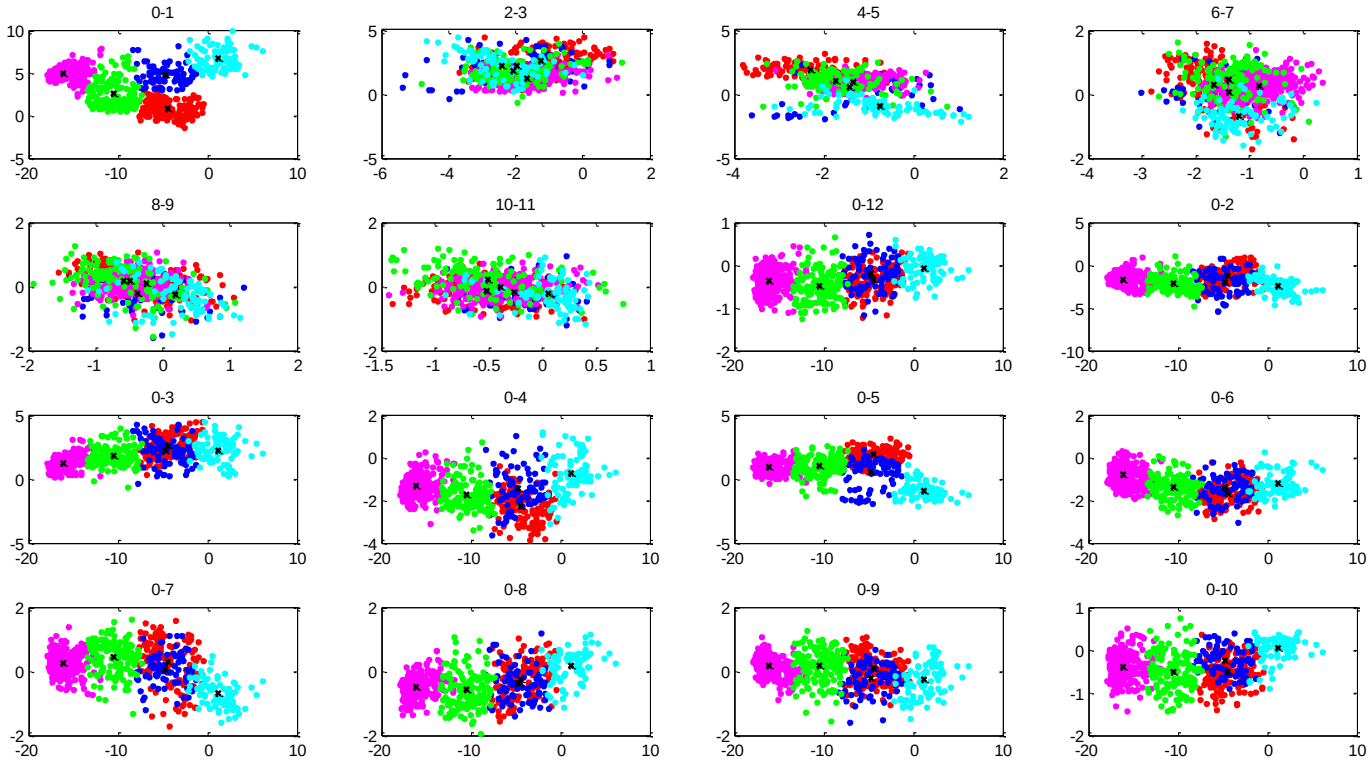
Σχήμα 13 : κίνηση των πλαισίων στον χώρο των τάξεων για 5 ομάδες (ψηφίο "4")

Στα παραπάνω διαγράμματα μπορούμε να παρατηρήσουμε μια κοινή μορφή που κατα κύριο λόγο προκύπτει απο την δεύτερη σειρά του σχήματος. Ωστόσο σε καμία περίπτωση δεν μπορούμε να πούμε οτι υπάρχει ένα συγκεκριμένο μοντέλο που μπορεί να εκφράσει με ακρίβεια όλες τις εκφωνήσεις του ψηφίου αυτού. Για μικρότερο αριθμό κλάσεων το μοντέλο είναι αρκετά "φτωχό" ενω για μεγάλο πλήθος κλάσεων δεν μπόρεσαν να παρατηρηθούν μεγάλες ομοιότητες όπως παραπάνω. Στο επόμενο σχήμα φαίνεται η κίνηση των πλαισίων για την εκφώνηση του ίδιου ψηφίου, για δυο διαφορετικές τιμές κλάσεων.

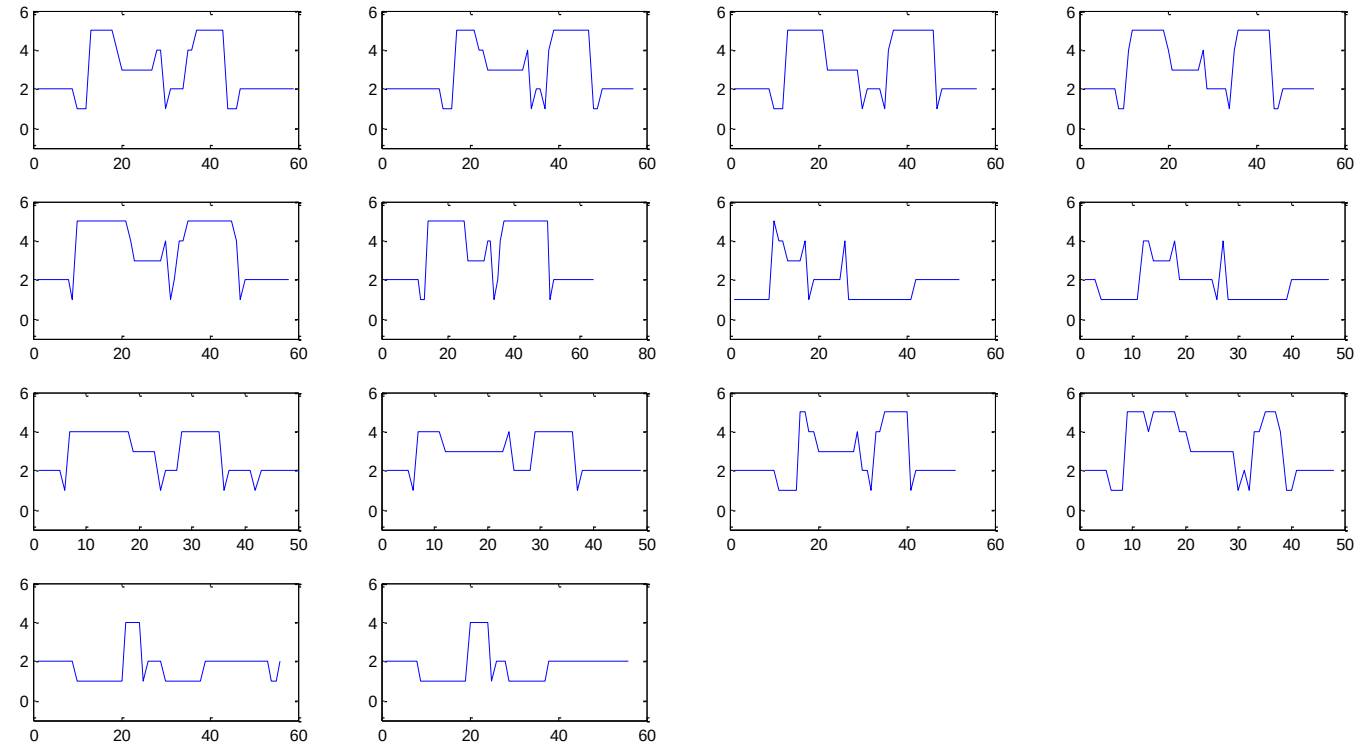


Σχήμα 14 : κίνηση πλαισίων στον χώρο τάξεων για 2 ομάδες (αριστερά) και για 8 ομάδες(δεξιά)

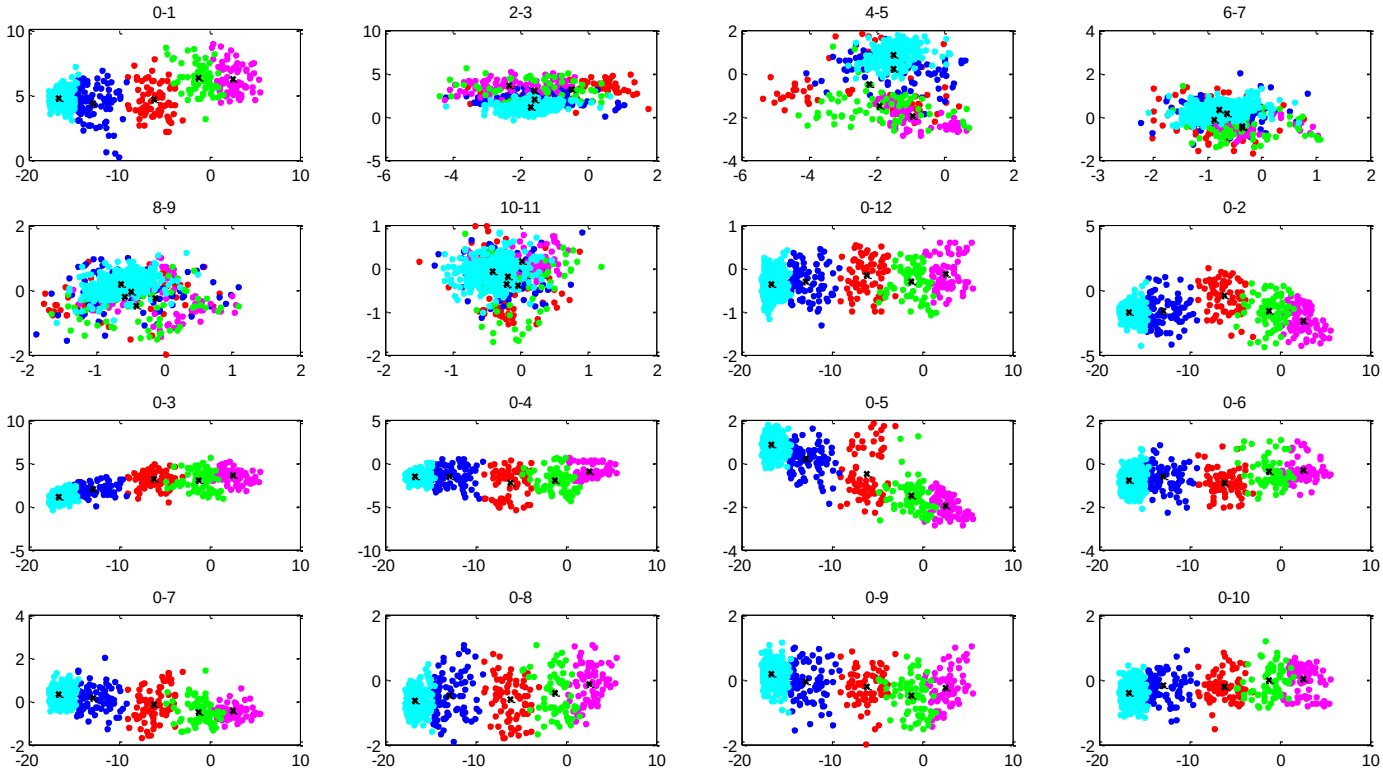
Κατα αντιστοιχία με τα παραπάνω η διαδικασία επαναλήφθηκε για το ψηφίο "6" και το ψηφίο "8". Σε καμία περίπτωση ωστόσο δεν παρατηρήθηκε ακριβής ομοιότητα κατα την κίνηση των πλαισίων για κάθε ψηφίο. Βέβαια μπορούμε και πάλι να παρατηρήσουμε όμοιες δομές στις γραφικές παραστάσεις οι οποίες κατα κύριο λόγο παρατηρήθηκαν για 5 clusters. Για λόγους πληρότητας παρουσιάζονται τα παρακάτω διαγράμματα των εκφωνήσεων αυτών.



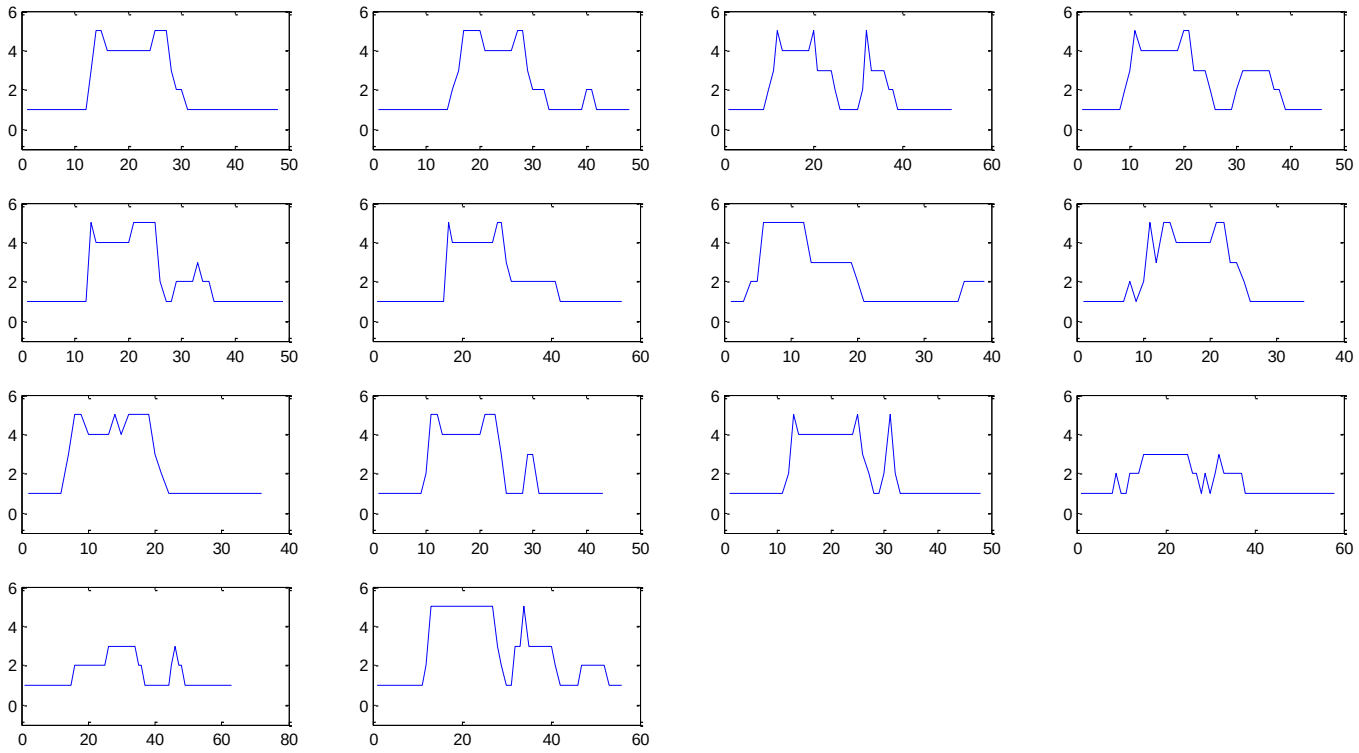
Σχήμα 15: απεικόνιση των clusters μετά απο ομαδοποίηση με kmeans (ψηφίο "6")



Σχήμα 16 : κίνηση των πλαισίων στον χώρο των τάξεων για 5 ομάδες (ψηφίο "6")



Σχήμα 17: απεικόνιση των clusters μετά απο ομαδοποίηση με kmeans (ψηφίο "8")



Σχήμα 18 : κίνηση των πλαισίων στον χώρο των τάξεων για 5 ομάδες (ψηφίο "8")

Ο λόγος για τον οποίο γίνεται μια τέτοια αναζήτηση, είναι η εύρεση ενός καθολικού μοντέλου το οποίο θα εκφράζει κάθε ψηφίο ξεχωριστά. Έτσι θα μπορεί να δίνεται μια νέα εκφώνηση ενός τυχαίου ψηφίου και θα μπορούμε να αποφανθούμε για ποιό ψηφίο πρόκειται παρατηρώντας την κλάση των παραθυρωμένων χρονικών πλαισίων του και συγκρίνοντάς την με αυτή του κάθε μοντέλου. Έτσι άν έχουμε ένα μοντέλο για κάθε ψηφίο, τότε το ψηφίο το οποίο εκφωνήθηκε θα είναι αυτό του οποίου το μοντέλο παρουσιάζει την πιο κοινή κίνηση των πλαισίων στον χώρο των τάξεων με αυτήν της νέας εκφώνησης. Στα προηγούμενα ερωτήματα είδαμε οτι οι MFCCs μεγαλύτερης τάξης δεν είναι ικανοί για έναν διαχωρισμό των πλαισίων σε αξιόπιστες τάξεις, ενώ αναφέραμε πως ο πρώτος και ο δεύτερος συντελεστής είναι οι σημαντικότεροι, κάτι που φάνηκε και απο την προβολή του αποτελέσματος του kmeans στα 2D επίπεδα ζευγών συντελεστών. Ακόμα και με όλους τους συντελεστές βέβαια ο αλγόριθμος kmeans δεν είναι αρκετός ωστε να μας παράξει ένα εύρωστο μοντέλο που θα μπορεί να εφαρμοστεί για κάθε ομιλητή καθώς οι ιδιαιτερότητες που μπορεί να παρουσιαστούν στον τρόπο ομιλίας κάθε ατόμου δεν μπορούν να μοντελοποιηθούν επαρκώς. Το πιο ορθό λοιπόν θα ήταν να μην αντιστοιχούμε μια κλάση σε κάθε παραθυροποιημένο πλαίσιο, αλλα την πιθανότητα να ανήκει στην κλάση αυτή, όπως και τις πιθανότητες να ανήκει στις άλλες κλάσεις. Κάτι τέτοιο αναμένεται να δούμε στα επόμενα εργαστηριακά θέματα με χρήση HMMs. Όλα όσα περιγράφηκαν στο 5ο μέρος υλοποιείται απο το παραδωταίο script **meros5.m**