ΕΡΓΑΣΤΗΡΙΑΚΗ 5 ΑΣΚΗΣΗ



# Αυτόματη Αναγνώριση Ομιλητή

## 1. Εισαγωγή

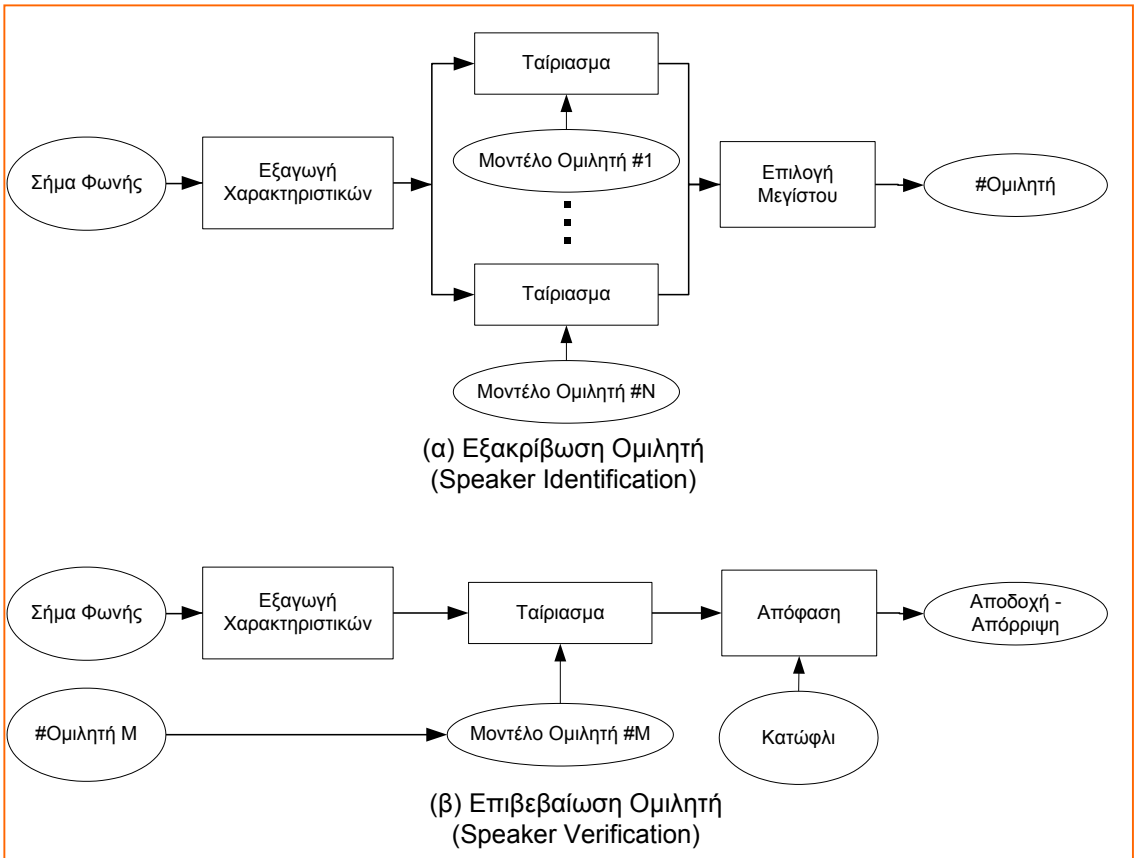
Με τον όρο «*αυτόματη αναγνώριση ομιλητή*» (automatic speaker recognition) εννοούμε την αυτόματη διαδικασία αναγνώρισης του προσώπου που μιλάει χρησιμοποιώντας την ιδιαίτερη πληροφορία που κρύβεται στην κυματομορφή της ομιλίας του. Με την τεχνική αυτή μπορούμε να χρησιμοποιήσουμε την φωνή προκειμένου να εξακριβώσουμε ή να επιβεβαιώσουμε την ταυτότητά ενός ομιλητή, και επομένως να του παρέχουμε πρόσβαση σε διάφορες υπηρεσίες, όπως για παράδειγμα πρόσβαση εισόδου σε χώρους ασφαλείας, παροχή υπηρεσιών μέσω διαδικτύου (web-banking), απομακρυσμένη πρόσβαση σε υπολογιστές κ.α..
Στόχος της συγκεκριμένης άσκησης είναι να δημιουργήσουμε ένα απλό αλλά ωστόσο αντιπροσωπευτικό σύστημα αυτόματης αναγνώρισης ομιλητή.

## 2. Αναγνώριση Ομιλητή

Η αναγνώριση ομιλητή (speaker recognition) μπορεί να διακριθεί σε *εξακρίβωση ομιλητή* (speaker identification) και *επιβεβαίωση ομιλητή* (speaker verification) [1], [2]. Κατά την εξακρίβωση ομιλητή, στόχος είναι να αναγνωρίσουμε ποιος ομιλητής έχει προφέρει μια φράση. Κατά την πιστοποίηση, ο ομιλητής προφέρει μια φράση και ταυτόχρονα δηλώνει την ταυτότητά του, ενώ στόχος του συστήματος είναι να εξακριβώσει αν πραγματικά είναι αυτός που δηλώνει. Στο Σχήμα 1 φαίνεται η βασική δομή των δύο συστημάτων.
Η αναγνώριση ομιλητή μπορεί επίσης να διακριθεί σε μεθόδους *εξαρτημένες και ανεξάρτητες από κείμενο* (text dependent/ independent). Σε ένα σύστημα ανεξάρτητο από κείμενο, το μοντέλο ομιλητή περιλαμβάνει χαρακτηριστικά του ομιλητή που εμφανίζονται ανεξάρτητα από τις φράσεις που προφέρει. Αντίθετα, σε ένα σύστημα εξαρτημένο από κείμενο, η αναγνώριση του ομιλητή βασίζεται σε συγκεκριμένες φράσεις που καλείται να προφέρει ο ομιλητής, όπως passwords, αριθμούς, PINs, κ.λ.π.
Γενικά, κάθε σύστημα αναγνώρισης ομιλητή περιλαμβάνει δύο βασικά τμήματα:

- την *εξαγωγή χαρακτηριστικών* (feature extraction), και
- την *ταυτοποίηση χαρακτηριστικών* (feature matching).

Εξαγωγή χαρακτηριστικών είναι η διαδικασία κατά την οποία εξάγεται περιορισμένη πληροφορία από το σήμα φωνής, η οποία όμως είναι ικανή να αναπαραστήσει κάθε ομιλητή.




Σχήμα 1: Συστήματα Αναγνώρισης Ομιλητή.

Η ταυτοποίηση των χαρακτηριστικών είναι ουσιαστικά η διαδικασία αναγνώρισης ενός άγνωστου ομιλητή συγκρίνοντας τα φωνητικά χαρακτηριστικά του με αυτά όλων των ομιλητών.

Κάθε σύστημα αναγνώρισης ομιλητή λειτουργεί σε δύο φάσεις. Η πρώτη είναι η *φάση της εγγραφής ή φάση εκπαίδευσης* (enrolment/ training phase), και η δεύτερη είναι η *φάση της λειτουργίας ή ελέγχου* (operation / testing phase). Κατά την εγγραφή, κάθε χρήστης του συστήματος παρέχει έναν αριθμό δειγμάτων ηχητικού σήματος, με τα οποία το σύστημα θα δημιουργήσει – εκπαιδεύσει το μοντέλο αναφοράς του συγκεκριμένου ομιλητή. Κατά τη διαδικασία λειτουργίας, το ηχητικό σήμα εισόδου αντιστοιχίζεται σε ένα από τα υπάρχοντα μοντέλα ομιλητή και πραγματοποιείται η αναγνώριση.

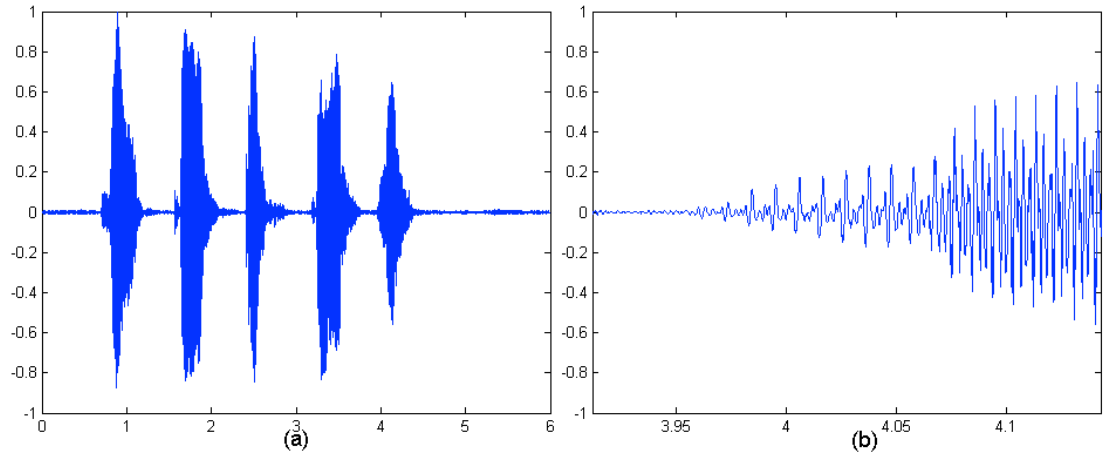
Η αναγνώριση ομιλητή είναι ένας δύσκολος στόχος και αποτελεί ένα ανοιχτό ερευνητικό πεδίο. Η λειτουργία ενός τέτοιου συστήματος βασίζεται στην ιδέα ότι το φωνητικό σήμα διαθέτει χαρακτηριστικά που είναι διακριτικά για κάθε ομιλητή. Κάτι τέτοιο ωστόσο δεν είναι εύκολο να ισχύει πάντοτε λόγω της τυχαιότητας των φωνητικών σημάτων. Τα φωνητικά σήματα κατά την φάση εκπαίδευσης και λειτουργίας μπορούν να διαφέρουν σημαντικά επειδή η φωνή ενός χρήστη μπορεί να αλλάξει με το χρόνο, τις συνθήκες υγείας, το ρυθμό ομιλίας, τη συναισθηματική κατάσταση, κ.ά. Επίσης, παράγοντες όπως ο διαφορετικός εξοπλισμός (π.χ. μικρόφωνα), το περιβάλλον ηχογράφησης, και η ύπαρξη ακουστικού θορύβου, μπορούν να επηρεάσουν σημαντικά την απόδοση ενός συστήματος αναγνώρισης ομιλητή.

# 3. Εξαγωγή Φωνητικών Χαρακτηριστικών

## 3.1. Εισαγωγή

Σκοπός του τμήματος αυτού είναι η μετατροπή της φωνητικής κυματομορφής σε κάποιο παραμετρικό τύπο αναπαράστασης με σημαντικά χαμηλότερο ρυθμό πληροφορίας για περαιτέρω ανάλυση και επεξεργασία.

Το σήμα φωνής μεταβάλλεται χρονικά με αργό τρόπο, και για αυτό χαρακτηρίζεται ως σχεδόν-στάσιμο (quasi-stationary). Ένα παράδειγμα ηχητικού σήματος φαίνεται στο Σχήμα 2. Στο Σχήμα 2.(α), ένας ομιλητής προφέρει στα αγγλικά την ακολουθία αριθμών «0-2-8-5-9» μέσα σε 6sec. Παρατηρούμε ότι το ηχητικό σήμα αλλάζει ταχύτατα (περίπου ανά 1/5sec), εφόσον αντιστοιχεί στους διαφορετικούς ήχους που προφέρονται. Στο Σχήμα 2.(β) επικεντρώνουμε το γράφημα σε μια περιοχή 0,25sec (αντιστοιχεί στην έναρξη της λέξης «nine»). Εκεί, διαπιστώνουμε ότι σε μικρά χρονικά τμήματα της τάξης των 5-100msec, τα χαρακτηριστικά του σήματος είναι περίπου σταθερά. Επομένως, ένας αποδοτικός τρόπος αναπαράστασης του φωνητικού σήματος είναι η φασματική ανάλυση περιορισμένου χρόνου (short-time spectral analysis) [3].




Σχήμα 2: Σήμα ομιλίας.

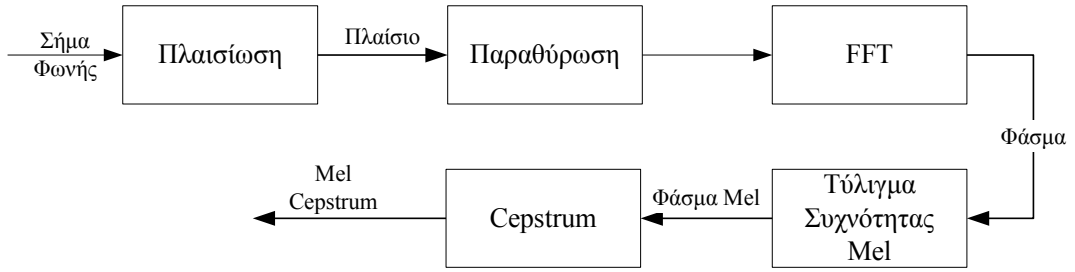
Υπάρχει μια μεγάλη ποικιλία χαρακτηριστικών και μεθόδων που μπορούν να χρησιμοποιηθούν προκειμένου να αναπαραστήσουμε παραμετρικά το σήμα φωνής για αναγνώριση ομιλητή: όπως οι συντελεστές γραμμικής πρόβλεψης (Linear Prediction Coefficients, LPC), οι συντελεστές συχνότητας Cepstrum του Mel (Mel-Frequency Cepstrum Coefficients, MFCC), και άλλοι [2], [4], [5].

Οι συντελεστές MFCC είναι ίσως οι πιο γνωστές παράμετροι για αναγνώριση ομιλητή, και αυτές θα χρησιμοποιήσουμε στην παρούσα άσκηση. Βασίζονται στο φασματικό περιεχόμενο του φωνητικού σήματος, αλλά όπως αυτό αναγνωρίζεται από την ανθρώπινη ακοή. Συγκεκριμένα, εφαρμόζουμε τριγωνικά φίλτρα των οποίων οι κεντρικές συχνότητες ισαπέχουν στις χαμηλές συχνότητες και λογαριθμικά τοποθετημένα στις υψηλότερες συχνότητες. Η διαδικασία αυτή ορίζει την κλίμακα Mel, και ορίζει γραμμικές αποστάσεις κάτω από τα 1000Hz, και λογαριθμικές αποστάσεις πάνω από τα 1000Hz. Στην επόμενη παράγραφο περιγράφεται η διαδικασία εξαγωγής των συντελεστών MFCC.

## 3.2. Εξαγωγή Συντελεστών MFCC

Στο επόμενο σχήμα περιγράφεται η διαδικασία εξαγωγής των MFCC. Αρχικά, το ηχητικό σήμα ηχογραφείται με ρυθμό δειγματοληψίας πάνω από 8KHz. Το δειγματοληπτημένο σήμα περιέχει

συχνότητες μέχρι τα 4KHz, οι οποίες καλύπτουν τη μεγαλύτερη ενέργεια των ήχων που παράγονται από τους ανθρώπους. Όπως αναφέρθηκε παραπάνω, κατά την εξαγωγή των MFCC προσπαθούμε να μιμηθούμε την ανθρώπινη ακοή. Οι συντελεστές MFCC, σε αντίθεση με το ίδιο το ηχητικό σήμα, είναι λιγότερο ευαίσθητοι στις διακυμάνσεις που αναφέραμε.




Σχήμα 3: Διαδικασία Εξαγωγής Συντελεστών MFCC.

Στα πλαίσια της άσκησης, θα εργαστείτε με ηχητικά σήματα που έχουν ήδη ηχογραφηθεί (δειγματοληπτηθεί και κβαντιστεί) και σας δίνονται υπό τη μορφή αρχείων wav.[1]

### *Πλαισίωση (Framing)*

Στο βήμα αυτό, η συνεχής ροή δειγμάτων ηχητικού σήματος διαιρείται σε πλαίσια των N δειγμάτων, ενώ τα διαδοχικά πλαίσια απέχουν μεταξύ τους κατά M δείγματα (M<N). Στην άσκηση αυτή, θα χρησιμοποιήσουμε N=256 και M=100. Αν το ηχητικό σήμα έχει δειγματοληπτηθεί με $f_s$=8KHz, τότε η περίοδος δείγματος είναι $T_s=1/f_s$=0,125msec. Έτσι, το σήμα διαιρείται σε πλαίσια διάρκειας $NT_s$=32msec, τα οποία απέχουν κατά $MT_s$=12,5msec. Αυτό σημαίνει ότι εξάγεται ένα πλαίσιο κάθε 12,5msec, και ότι τα πλαίσια είναι επικαλυπτόμενα.

Σε κώδικα MATLAB, αν n είναι ο δείκτης κάθε πλαισίου, τότε το πλαίσιο περιλαμβάνει τα δείγματα του σήματος (n-1)*M+1:(n-1)*M+N. Έτσι, ο αριθμός των πλαισίων που θα προκύψουν από ένα σήμα T δειγμάτων θα είναι $B = \left\lfloor \frac{T-N}{M} + 1 \right\rfloor$, υποθέτοντας ότι «πετάμε» τελευταία δείγματα του σήματος.

### *Παραθύρωση*

Το επόμενο βήμα είναι η εφαρμογή ενός παραθύρου σε κάθε πλαίσιο προκειμένου να ελαχιστοποιήσουμε τις ασυνέχειες του σήματος στην αρχή και το τέλος του πλαισίου. Η βασική ιδέα είναι επιβάλλουμε ένα «σβήσιμο» του σήματος στα άκρα του κάθε πλαισίου. Εάν ορίσουμε ένα παράθυρο ως w(n), n=0:N-1, όπου N είναι ο αριθμός δειγμάτων ανά πλαίσιο, τότε η παραθύρωση του πλαισίου l δίνει το σήμα

$$y_l(n) = x_l(n)w(n), \quad 0 \le n \le N-1.$$

Συνήθως, χρησιμοποιείται το παράθυρο Hamming που ορίζεται ως:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1.$$

[1] Για να διαβάσετε ένα αρχείο wav εκτελείται στο Matlab την εντολή [x,fs,Nbits] = wavread('filename.wav'), όπου x είναι το ηχητικό σήμα μήκους M δειγμάτων, δειγματοληπτημένο με συχνότητα fs και κβαντισμένο με Nbits. Σημειώστε ότι οι τιμές του σήματος x σε ένα αρχείο wav έχουν δυναμική περιοχή [-1:1]. Τα αρχεία που σας δίνονται χρησιμοποιούν 16 bits, και έχουν δειγματοληπτηθεί με fs=8KHz, άρα είναι διάρκειας M*0,125msecs.

### *Fast Fourier Transform (FFT)*

Το επόμενο βήμα επεξεργασίας είναι ο FFT, που δέχεται ένα πλαίσιο N δειγμάτων και το μεταφέρει από το πεδίο του χρόνου στο πεδίο της συχνότητας. Ο αλγόριθμος FFT είναι ένας αποδοτικός αλγόριθμος για την υλοποίηση του Διακριτού Μετασχηματισμού Fourier (DFT). Αν σε ένα πλαίσιο που βρίσκεται αποθηκευμένο στο διάνυσμα x, εφαρμόσω FFT xf=fft(x);
τότε προκύπτουν N μιγαδικοί αριθμοί που ερμηνεύονται ως εξής:

- το xf(1) αντιστοιχεί στη μηδενική συχνότητα (DC),
- οι θετικές συχνότητες $0<f<f_s/2$ αντιστοιχούν στις τιμές για n=[2:N/2],
- οι αρνητικές συχνότητες $-f_s/2<f<0$ προκύπτουν για n=[N/2+1:N].

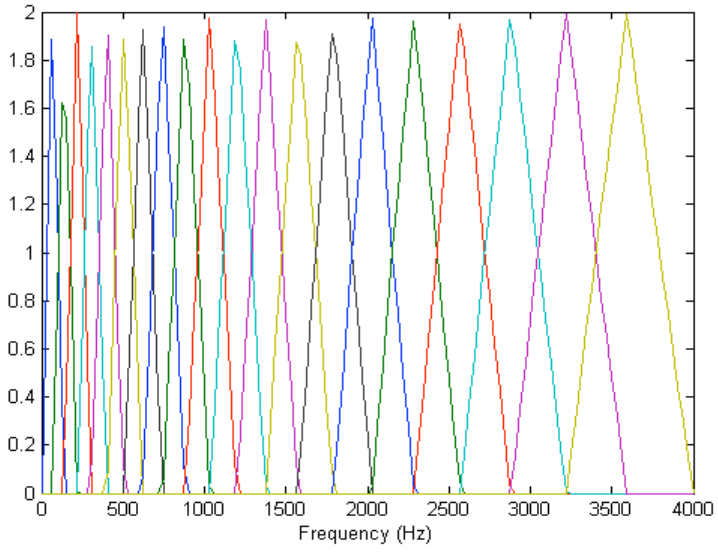
Το τετράγωνο του μέτρου του xf καλείται «φάσμα του σήματος» ή «περιοδόγραμμα». Δεδομένης της συμμετρίας του φάσματος στις αρνητικές και θετικές συχνότητες, αρκεί να περιοριστούμε στις θετικές συχνότητες. Άρα, το φάσμα ενός πλαισίου δίνεται ως sp = abs(xf(1:N/2+1)).^2;

### *Τύλιγμα Συχνότητας Mel (Mel-frequency Wrapping)*

Ψυχοακουστικές έρευνες έχουν δείξει ότι η ανθρώπινη αντίληψη για τις συχνότητες ενός φωνητικού σήματος δεν ακολουθεί γραμμική κλίμακα. Έτσι, για κάθε τόνο με πραγματική συχνότητα f μετρούμενη σε Hz, υπάρχει ένας υποκειμενικός τόνος στην κλίμακα mel. Η κλίμακα συχνοτήτων mel είναι γραμμική μέχρι το 1KHz, και λογαριθμική στις υψηλότερες συχνότητες. Ως σημείο αναφοράς, ο τόνος των 1KHz (40dB πάνω από το κατώφλι ακουστότητας) ορίζεται ως 1000 mels. Επομένως, μπορούμε να χρησιμοποιήσουμε την ακόλουθη προσεγγιστική σχέση για να υπολογίσουμε σε mels μια συχνότητα f που δίνεται σε Hz:

$$mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right).$$

Μια προσέγγιση για να προσομοιώσουμε το υποκειμενικό φάσμα είναι να χρησιμοποιήσουμε μια διάταξη φίλτρων: ένα φίλτρο για κάθε επιθυμητή μελ-συχνότητα, όπως φαίνεται στο παρακάτω σχήμα. Αυτή η διάταξη των φίλτρων έχει τριγωνική ζωνοδιαβατή απόκριση συχνότητας, και η απόσταση καθώς και το εύρος ζώνης κάθε φίλτρου καθορίζεται από ένα σταθερό διάστημα μελ-συχνότητας. Το τροποποιημένο φάσμα του S(ω), έστω $S_m(\omega)$, αποτελείται πλέον από την ισχύ του σήματος στην έξοδο κάθε φίλτρου αν στην είσοδό του εφαρμοστεί το S(ω). Ο αριθμός των μελ συντελεστών φάσματος συνήθως επιλέγεται K=20.




Σχήμα 4: Διάταξη Φίλτρων τοποθετημένων κατά την κλίμακα του Mel.

Παρατηρήστε ότι η διάταξη των φίλτρων εφαρμόζεται απευθείας στο πεδίο της συχνότητας. Επομένως, για να πάρουμε το φάσμα Mel, αρκεί να πολλαπλασιάσουμε τα τριγωνικά παράθυρα με το φάσμα του σήματος ομιλιας.
Ο σχεδιασμός των τριγωνικών παραθύρων δίνεται έτοιμος ως ένας πίνακας MEL, διάστασης 20x129 τιμών. Κάθε γραμμή του πίνακα αντιστοιχεί σε ένα τριγωνικό παράθυρο. Προκειμένου να απεικονίσετε τα παράθυρα, εκτελείται:

```
load mel.mat
plot(linspace(0, 4000, 129), MEL');
```

Για να υπολογίσετε το φάσμα ενός πλαισίου σε κλίμακα Mel, αρκεί να εκτελέσετε: msp=MEL*sp;
οπότε το διάνυσμα msp περιλαμβάνει 20 συντελεστές του μελ φάσματος του πλαισίου.

### *Cepstrum*[2]

Το τελικό βήμα εξαγωγής των χαρακτηριστικών φωνής είναι ο υπολογισμός του Μελ-Cepstrum. Συγκεκριμένα, υπολογίζουμε το φυσικό λογάριθμο του μελ φάσματος, και κατόπιν το επιστρέφουμε στο πεδίο του χρόνου. Το αποτέλεσμα της διαδικασίας αυτής είναι οι συντελεστές MFCC. Η αναπαράσταση του φασματικού περιεχομένου ενός σήματος μέσω του cepstrum δίνει μια καλή περιγραφή των φασματικών ιδιοτήτων του σήματος σε ένα περιορισμένο χρονικά παράθυρο ανάλυσης.
Επειδή οι συντελεστές μελ φάσματος (και ο λογάριθμός τους) είναι πραγματικοί αριθμοί, μπορούμε να τους γυρίσουμε στο πεδίο του χρόνου χρησιμοποιώντας το Διακριτό Μετασχηματισμό Συνημιτόνου (Discrete Cosine Transform, DCT):

$$c(n) = \sum_{k=1}^{K} \ln\left(S_m(k)\right) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \quad n = 1,\ldots,K$$

Άρα, οι συντελεστές MFCC για ένα πλαίσιο φωνητικού σήματος δίνονται ως
mfcc=dct(log(msp));
Σημειώστε ότι συνήθως εξαιρείται ο πρώτος συντελεστής MFCC, $c_0$, από το αποτέλεσμα του DCT. Ο συντελεστής αυτός ουσιαστικά αναπαριστά τη μέση τιμή του σήματος εισόδου και άρα δε φέρει σημαντική πληροφορία.
Μέσω της διαδικασίας που περιγράφηκε παραπάνω, κάθε πλαίσιο του φωνητικού σήματος εισόδου N=256 δειγμάτων (32msec) μετατρέπεται σε ένα διάνυσμα K-1=19 συντελεστών MFCC, που αποτελεί μια συνοπτική και περιεκτική αναπαράσταση του συχνοτικού περιεχομένου του πλαισίου. Τα διανύσματα των MFCC λέγονται και ακουστικά διανύσματα και στην άσκηση αυτή εξάγονται με ρυθμό 1/12,5msec. Στο τέλος της διαδικασίας εξαγωγής χαρακτηριστικών, το φωνητικό σήμα ενός ομιλητή έχει μετατραπεί σε μια ακολουθία ακουστικών διανυσμάτων. Στην επόμενη ενότητα θα δούμε πώς χρησιμοποιούνται τα ακουστικά διανύσματα για την αναγνώριση ομιλητή.

**Ερώτημα 1:** Ακολουθώντας τις οδηγίες που περιγράφηκαν στην ενότητα αυτή, υλοποιήστε μια συνάρτηση στη MATLAB που να υλοποιεί το τμήμα εξαγωγής των συντελεστών MFCC. Η συνάρτηση θα περιγράφεται ως mfcc = mfcc_extractor(x);
όπου το διάνυσμα x θα περιλαμβάνει το φωνητικό σήμα μιας ηχογράφησης, διάρκειας T δειγμάτων, και ο πίνακας mfcc, διάστασης (K-1)xB, περιλαμβάνει την ακολουθία των ακουστικών διανυσμάτων. Συγκεκριμένα, στην b στήλη του περιλαμβάνεται το (K-1)x1 διάνυσμα των MFCC του b-οστού πλαισίου.

---

[2] Η λέξη Cepstrum είναι αναγραμματισμός του Spectrum.

# 4. Αντιστοίχιση Χαρακτηριστικών (Feature Matching)

## 4.1. Εισαγωγή

Το πρόβλημα της αναγνώρισης ομιλητή ανήκει σε μια ευρύτερη επιστημονική περιοχή που καλείται αναγνώριση προτύπων (pattern recognition). Στόχος της αναγνώρισης προτύπων είναι η ταξινόμηση αντικειμένων σε μία από περισσότερες κατηγορίες. Τα αντικείμενα που θέλουμε να ταξινομήσουμε λέγονται πρότυπα (patterns). Στην περίπτωσή μας, οι ακολουθίες των ακουστικών διανυσμάτων αποτελούν τα πρότυπα, ενώ οι ομιλητές αποτελούν τις κατηγορίες. Επειδή η διαδικασία ταξινόμησης εφαρμόζεται σε εξαγόμενα χαρακτηριστικά, μπορεί να ονομαστεί και ταίριασμα χαρακτηριστικών (feature matching).

Στην περίπτωση που διαθέτουμε κάποια πρότυπα και γνωρίζουμε την κατηγορία που ανήκει το καθένα, τότε μιλάμε για επιβλεπόμενη αναγνώριση προτύπων. Ουσιαστικά αυτό γίνεται κατά τη φάση της εκπαίδευσης του συστήματος αναγνώρισης ομιλητή: διαθέτουμε τα ακουστικά διανύσματα και γνωρίζουμε ότι εξάχθηκαν από συγκεκριμένο ομιλητή. Τα πρότυπα αυτά αποτελούν το σύνολο εκπαίδευσης. Τα υπόλοιπα πρότυπα χρησιμοποιούνται για να ελέγξουν τον αλγόριθμο ταξινόμησης, και αποκαλούνται σύνολο ελέγχου. Εάν κατά τη φάση του ελέγχου, γνωρίζουμε τις κλάσεις των προτύπων ελέγχου, τότε μπορούμε να εκτιμήσουμε την απόδοση του συστήματος.

Οι τεχνικές ταυτοποίησης χαρακτηριστικών που χρησιμοποιούνται σε εφαρμογές αναγνώρισης ομιλητή είναι οι Dynamic Time Warping (DTW), Hidden Markov Models (HMM), και Vector Quantization (VQ) [1], [2]. Στα πλαίσια της άσκησης, θα υλοποιήσουμε τη μέθοδο της διανυσματικής κβάντισης (VQ), λόγω της απλότητας και της σχετικά καλής ακρίβειας που προσφέρει.

## 4.2. Διανυσματική Κβάντιση (VQ)

Η διαδικασία της διανυσματικής κβάντισης έχει ποικίλες εφαρμογές πέρα από την αναγνώριση προτύπων. Είναι ουσιαστικά μια γενίκευση της διαδικασίας της μη ομοιόμορφης κβάντισης από βαθμωτούς αριθμούς σε διανύσματα. Ένας διανυσματικός κβαντιστής δέχεται ως είσοδο διανύσματα μεγέθους Kx1 που έχουν διάφορες τιμές, και επιστρέφει κωδικές λέξεις (codewords) μεγέθους Kx1, οι οποίες έχουν συγκεκριμένες τιμές. Το σύνολο των κωδικών λέξεων λέγεται βιβλίο κωδικών λέξεων (codebook), και έχει πεπερασμένο μέγεθος C. Τα διανύσματα εισόδου του κβαντιστή ανήκουν σε ένα Kx1 διανυσματικό χώρο, τον οποίο ο διανυσματικός κβαντιστής αναπαριστά μέσω C περιοχών (clusters). Κάθε περιοχή αναπαρίσταται από το κέντρο μάζας της (centroid) που είναι η αντίστοιχη κωδική λέξη.

Έστω ότι έχουμε το βιβλίο των C κωδικών λέξεων, και θέλουμε να κβαντίσουμε ένα διάνυσμα **x**. Ο διανυσματικός κβαντιστής υπολογίζει την απόσταση ανάμεσα στο διάνυσμα εισόδου και κάθε κωδική λέξη, $\mathbf{c}_n$, n=1,…,C. Η κωδική λέξη που θα έχει τη μικρότερη απόσταση, είναι και η λέξη στην οποία θα κβαντιστεί το διάνυσμα εισόδου:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{c}_n \in \mathcal{E}} d\left(\mathbf{x}, \mathbf{c}_n\right)$$

Μέχρι στιγμής δεν έχουμε ορίσει την έννοια της απόστασης d. Η παράμετρος αυτή εξαρτάται από την εφαρμογή, και υπάρχουν διάφορες επιλογές. Στην περίπτωσή μας, επιλέγουμε να χρησιμοποιήσουμε την πιο εύκολη μετρική που είναι το τετράγωνο της ευκλείδειας απόστασης των δύο διανυσμάτων [1]:

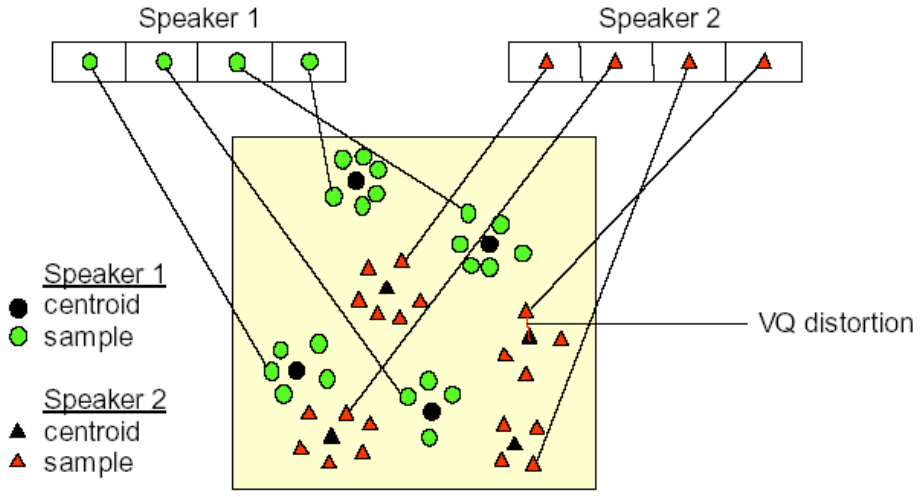
$$d\left(\mathbf{x}, \mathbf{c}_n\right) = \left|\mathbf{x} - \mathbf{c}_n\right|^2$$

που στο Matlab αντιστοιχεί στην εντολή norm(x-cn)^2.

Κατά τη διαδικασία σχεδιασμού του κβαντιστή, θέλουμε να προσδιορίσουμε τις κωδικές λέξεις του βιβλίου, ώστε να ελαχιστοποιήσουν την παραμόρφωση ανάμεσα στην είσοδο και την κβαντισμένη τιμή της. Αυτό σημαίνει ότι κάθε κωδική λέξη θα πρέπει να είναι το κέντρο μάζας της αντίστοιχης συστάδας. Εφόσον χρησιμοποιούμε την ευκλείδεια απόσταση, το κέντρο μάζας ισοδυναμεί με το μέσο όρο. Άρα, αν γνωρίζουμε ότι ένα υποσύνολο των διανυσμάτων εισόδου ανήκει σε μία συστάδα, τότε η κωδική λέξη που έχει την ελάχιστη απόσταση από όλα τα διανύσματα της συστάδας ορίζεται ως ο μέσος όρος των διανυσμάτων.
Όμως, κατά τη φάση σχεδιασμού του διανυσματικού κβαντιστή συνήθως δε γνωρίζουμε την κατανομή των διανυσμάτων εισόδου, και άρα την κατηγοριοποίηση τους σε συστάδες. Έτσι, η εύρεση των περιοχών κβάντισης και άρα των κωδικών λέξεων γίνεται με μια επαναληπτική διαδικασία που θα περιγράψουμε παρακάτω.

## 4.3. Διανυσματική Κβάντιση για Αναγνώριση Ομιλητή

Στην περίπτωσή μας, τα διανύσματα εισόδου του διανυσματικού κβαντιστή είναι τα ακουστικά διανύσματα και ανήκουν σε ένα διανυσματικό χώρο διάστασης K=19. Τα διανύσματα αυτά θέλουμε να τα κβαντίσουμε σε C κωδικές λέξεις, και να προκύψει ένα βιβλίο C κωδικών για κάθε ομιλητή. Στην άσκηση αυτή, θα επιλέξουμε C=4, 8 και 16. Αφού σχεδιάσουμε ένα διανυσματικό κβαντιστή (στην ουσία ένα βιβλίο κωδικών) για κάθε ομιλητή, μπορούμε να αναγνωρίσουμε ποιος ομιλητής πρόφερε μια συγκεκριμένη φράση: αρκεί να κβαντίσουμε τη φράση αυτή χρησιμοποιώντας κάθε κβαντιστή και να υπολογίσουμε την παραμόρφωση που προκύπτει για τον καθένα. Ο ομιλητής που θα έχει τη μικρότερη παραμόρφωση θα είναι και ο σωστός.
Στο επόμενο διάγραμμα περιγράφεται αφαιρετικά η διαδικασία αναγνώρισης. Στο σχήμα περιλαμβάνονται μόνο δύο ομιλητές, υποθέτουμε ότι τα ακουστικά διανύσματα έχουν μόνο δύο συντελεστές και χρησιμοποιούμε C=4 κωδικές λέξεις. Οι κύκλοι αναφέρονται στα ακουστικά διανύσματα του ομιλητή 1, ενώ τα τρίγωνα στον ομιλητή 2. Στη φάση εκπαίδευσης, δημιουργούμε ένα βιβλίο κωδικών λέξεων για κάθε ομιλητή ομαδοποιώντας κατάλληλα τα ακουστικά του διανύσματα. Οι κωδικές λέξεις που προκύπτουν απεικονίζονται στο σχήμα με μαύρους κύκλους και μαύρα τρίγωνα, για τον ομιλητή 1 και 2, αντίστοιχα. Η απόσταση ενός διανύσματος από την κοντινότερη κωδική λέξη λέγεται παραμόρφωση VQ (VQ-distortion).




Σχήμα 5: Δημιουργία Κωδικών Λέξεων του Διανυσματικού Κβαντιστή.

## 4.4. Κατηγοριοποίηση των Διανυσμάτων Εκπαίδευσης

Κατά τη φάση της εγγραφής, τα ακουστικά διανύσματα αποτελούν το σύνολο διανυσμάτων εκπαίδευσης, με τα οποία καλούμαστε να σχεδιάσουμε το διανυσματικό κβαντιστή (βιβλίο κωδικών) για κάθε ομιλητή. Για να παράγουμε ένα σύνολο C κωδικών λέξεων (codewords) χρησιμοποιώντας B διανύσματα εκπαίδευσης μπορούμε να χρησιμοποιήσουμε τον αλγόριθμο LBG (Linde, Buzo, Gray). Ο αλγόριθμος αυτός αποτελεί μια τροποποίηση του περισσότερο γνωστού αλγόριθμου Lloyd ή K-means [1], και υλοποιείται με την ακόλουθη επαναληπτική διαδικασία:

| **Αλγόριθμος LBG** | |
|---|---|
| Βήμα 1: | Δημιούργησε ένα βιβλίο με μία κωδική λέξη. Αυτή παράγεται ως ο μέσος όρος των διανυσμάτων εκπαίδευσης. |
| Βήμα 2: | Διπλασίασε το μέγεθος του τρέχοντος βιβλίου διαιρώντας κάθε κωδική λέξη σε δύο: $cw_n^+ = cw_n(1+\varepsilon)$ $cw_n^+ = cw_n(1-\varepsilon)$ όπου το n κυμαίνεται από 1 έως το μέγεθος του τρέχοντος βιβλίου, και ε είναι μια παράμετρος διαίρεσης (επιλέγουμε ε=0.01). |
| Βήμα 3: | Αναζήτηση Κοντινότερου Γείτονα. Για κάθε διάνυσμα εκπαίδευσης, βρες την κωδική λέξη του τρέχοντος βιβλίου στην οποία είναι πιο κοντά, και ανάθεσε το διάνυσμα εκπαίδευσης στη συγκεκριμένη κωδική λέξη. |
| Βήμα 4: | Ανανέωση των Μέσων. Ανανέωσε τα κέντρα κάθε κωδικής λέξης χρησιμοποιώντας το μέσο όρο των διανυσμάτων εκπαίδευσης που ανατέθηκαν σε αυτήν την περιοχή. |
| Βήμα 5: | Επανέλαβε τα βήματα 3 και 4 μέχρι η μέση απόσταση να είναι κάτω από ένα προκαθορισμένο κατώφλι. |
| Βήμα 6: | Επανέλαβε τα βήματα 2, 3 και 4, μέχρι το μέγεθος του βιβλίου να γίνει C. |

Ο αλγόριθμος LBG δημιουργεί ένα βιβλίο C κωδικών λέξεων σε επίπεδα. Ξεκινά με μία κωδική λέξη, και διαιρώντας τις κωδικές λέξεις αρχικοποιεί την αναζήτηση για το βιβλίο των 2 κωδικών λέξεων, κ.ο.κ. Στο παρακάτω σχήμα φαίνεται ένα διάγραμμα ροής του αλγορίθμου LBG.

**Ερώτημα 2:** Υλοποιήστε τον παραπάνω αλγόριθμο δημιουργίας βιβλίου κωδικών λέξεων. Συγκεκριμένα, υλοποιήστε μια ρουτίνα στη MATLAB με τα εξής χαρακτηριστικά:

Book = make_codebook(Features,C);

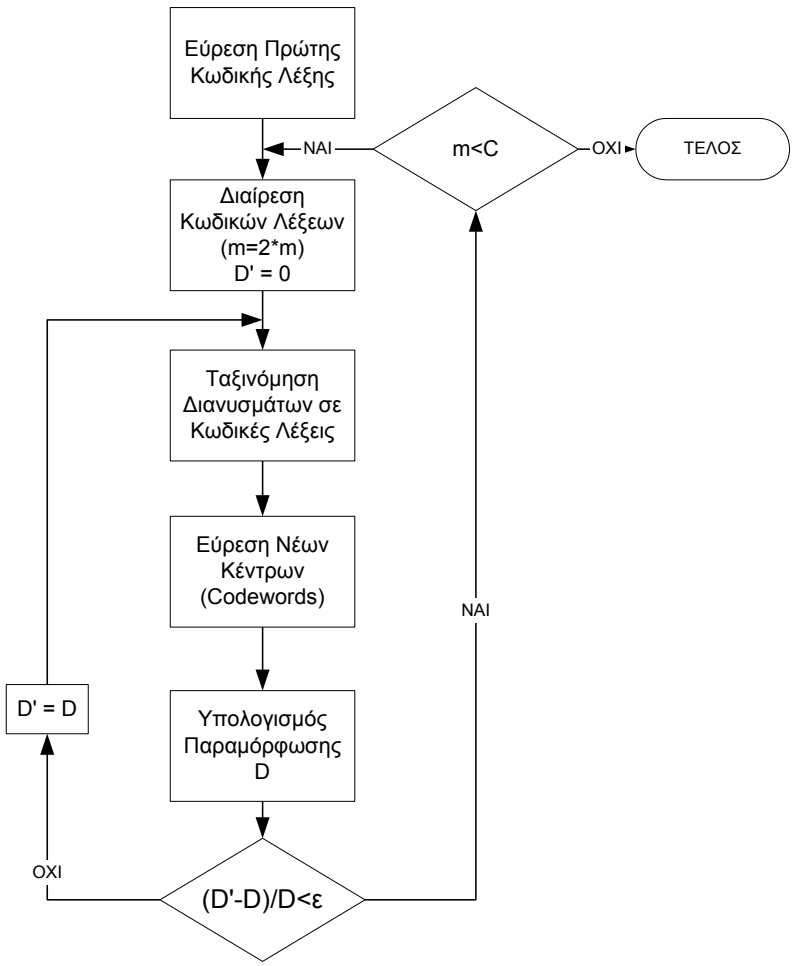
όπου Features θα είναι ένας πίνακας KxB και κάθε στήλη του περιέχει τους συντελεστές MFCC για ένα πλαίσιο ομιλίας ενός ομιλητή. C είναι ο αριθμός των κωδικών λέξεων που θα χρησιμοποιήσουμε. Book είναι ένας πίνακας KxC, και περιέχει τις C κωδικές λέξεις διάστασης Kx1 του ομιλητή αποθηκευμένες κατά στήλες.

Παρατηρήστε ότι η μέθοδος αναγνώρισης ομιλητή VQ δεν εμπεριέχει την έννοια της χρονικής ακολουθίας των πλαισίων. Έτσι, αν έχουμε υπολογίσει τους συντελεστές MFCC ενός ομιλητή από δύο φράσεις και τους έχουμε αποθηκεύσει στους πίνακες Features1, Features2, τότε ο κοινός κβαντιστής των δύο φράσεων του ομιλητή μπορεί να υπολογιστεί αν δώσουμε στην παραπάνω συνάρτηση

Features = [Features1 Features2];

Η συνάρτηση make_codebook θα πρέπει να καλεί τη συνάρτηση:

[Score,Distance,Index] = compute_distance(Features,Codes);




Σχήμα 6: Διάγραμμα Ροής Αλγορίθμου LBG [1].

που δέχεται ως είσοδο τα ακουστικά διανύσματα υπό τη μορφή του πίνακα Features (KxB), καθώς και τους κωδικούς του τρέχοντος βιβλίου στον πίνακα Codes ($KxC_i$). Παρατηρήστε ότι το μέγεθος του βιβλίου δεν είναι εξαρχής C, αλλά ξεκινά από 1 και διπλασιάζεται σε κάθε εξωτερικό βρόγχο του αλγορίθμου.

Οι έξοδοι της συνάρτησης είναι ο αριθμός Score, και τα Bx1 διανύσματα Distance και Index.

Η συνάρτηση αυτή υπολογίζει την απόσταση κάθε ακουστικού διανύσματος από κάθε κωδική λέξη του βιβλίου. Στη συνέχεια, για κάθε ακουστικό διάνυσμα επιλέγει τη μικρότερη απόσταση, την τοποθετεί στο διάνυσμα Distance, και αποθηκεύει το νούμερο της αντίστοιχης κωδικής λέξης στο διάνυσμα Index. Άρα, αν το 100-οστό ακουστικό διάνυσμα, Features(:,100), κβαντίζεται στην κωδική λέξη Codes(:,4) και έχει απόσταση από αυτήν 5,5, τότε αποθηκεύουμε Distance(100)=5.5, και Index(100)=4.

Ο αριθμός Score περιέχει το μέσο όρο των αποστάσεων για όλα τα διανύσματα εισόδου:

Score = mean(Distance);

Σημειώστε ότι η συνάρτηση compute_distance καλείται δύο φορές σε κάθε εσωτερικό βρόγχο του αλγορίθμου LBG:

- μια φορά για να ταξινομήσει τα ακουστικά διανύσματα στις περιοχές και να μας δώσει το Index προκειμένου να ανανεώσουμε τα κέντρα,
- και μια δεύτερη φορά, για να υπολογίσει την απόσταση των διανυσμάτων από τα ανανεωμένα κέντρα.

## 5. Ηχητικά Δεδομένα της Εφαρμογής

Για το σκοπό της άσκησης, δίνονται 32 αρχεία wav, διάρκειας 1sec (8000 δείγματα) το καθένα. Τα αρχεία αυτά περιλαμβάνουν ηχογραφήσεις από 8 άντρες ομιλητές που προφέρουν τη λέξη «one», τέσσερεις φορές ο καθένας. Τα αρχεία φέρουν το όνομα:
xy.wav
όπου το x αναφέρεται στον κωδικό του ομιλητή (x=1:8), και το y αναφέρεται στον αριθμό ηχογράφησης του συγκεκριμένου ομιλητή (y=1:4). Έτσι, το 62.wav περιέχει την δεύτερη ηχογράφηση του έκτου ομιλητή.
Οι ηχογραφήσεις με αριθμούς 1-2 θα χρησιμοποιηθούν για την εκπαίδευση του συστήματος, ενώ οι ηχογραφήσεις 3-4 θα χρησιμοποιηθούν για τον έλεγχο του συστήματος.

## 6. Εκπαίδευση του Συστήματος

**Ερώτημα 3:** Δημιουργήστε μια συνάρτηση train, η οποία **για κάθε ομιλητή** θα εκτελεί τα παρακάτω:

- άνοιγμα αρχείων x1.wav, x2.wav,
- υπολογισμός των MFCC συντελεστών για κάθε ηχογράφηση με χρήση της mfcc_extractor,
- συνένωση των MFCC συντελεστών από τις δύο ηχογραφήσεις,
- δημιουργία βιβλίου κωδικών για το συγκεκριμένο ομιλητή με χρήση της make_codebook,
- αποθήκευση του βιβλίου ως Bookx.

Επιλέξτε δύο συντελεστές MFCC, π.χ. δεύτερο και τρίτο, για δύο ομιλητές και σχεδιάστε τους στο ίδιο γράφημα με διαφορετικό χρώμα για κάθε ομιλητή. Εμφανίζονται κατά συστάδες; Μπορούν να διαχωριστούν οι συστάδες των δύο ομιλητών;

## 7. Έλεγχος του Συστήματος

**Ερώτημα 4:** Υλοποιήστε μια συνάρτηση test, η οποία για κάθε **ηχογράφηση** ελέγχου (x3.wav ,x4.wav), να εκτελεί τα παρακάτω:

- άνοιγμα ηχογράφησης,
- υπολογισμός των MFCC συντελεστών με χρήση της mfcc_extractor,
- υπολογισμός της παραμόρφωσης που επιτυγχάνεται χρησιμοποιώντας το βιβλίο κάθε ομιλητή με χρήση της συνάρτησης compute_distance, και του αριθμού Score,
- για την τρέχουσα ηχογράφηση, επιλέγεται ο ομιλητής με το μικρότερο Score.

Αφού εκτελέσετε την παραπάνω συνάρτηση, αναλύστε τα αποτελέσματα της αναγνώρισης. Ποιο είναι το ποσοστό της λανθασμένης ταξινόμησης; Τα λάθη εντοπίζονται μεταξύ συγκεκριμένων ομιλητών; Αν ναι, προσπαθήστε να τα δικαιολογήσετε μέσω της κατανομής των MFCC των προβληματικών ομιλητών.

**Ερώτημα 5:** Επαναλάβετε το προηγούμενο ερώτημα για C=4, 8 και 16 κωδικές λέξεις. Επίσης, δοκιμάστε να εκπαιδεύσετε το σύστημα χρησιμοποιώντας μόνο τις ηχογραφήσεις x1.wav, και ελέγξτε το με βάση τις υπόλοιπες (x2, x3, x4). Συμπληρώστε τον παρακάτω πίνακα ποσοστού σφαλμάτων, και εκτιμήστε πώς επηρεάζει η αύξηση των κωδικών λέξεων C, και το μέγεθος των δεδομένων εκπαίδευσης την απόδοση του συστήματος.

Ποια νομίζετε ότι θα ήταν η απόδοση του συστήματος εάν μειώναμε το πλήθος των συντελεστών MFCC;
Είναι λογικό να χρησιμοποιούμε όσο το δυνατόν μεγαλύτερο πλήθος παραμέτρων για να βελτιώσουμε την απόδοση ενός συστήματος αναγνώρισης ομιλητή;

| Απόδοση Συστήματος σε Ποσοστό Εσφαλμένων Αναγνωρίσεων (%) | | | |
|---|---|---|---|
| | Πλήθος Κωδικών Λέξεων (C) | | |
| Εκπαίδευση/Ελεγχος | 4 | 8 | 16 |
| 2-2 | 0,0625 | 0,0625 | 0,0000 |
| 1-3 | 0,1667 | 0,0833 | 0,0833 |

# 8. Βιβλιογραφία


[1] L.R. Rabiner, B.H. Juang, "Fundamentals of Speech Recognition", Prentice-Hall, Englewood Cliffs, N.J., 1993.
[2] J.P Campbell, "Speaker Recognition: A Tutorial", Proceedings of the IEEE, vol 85, No 9, September 1999.
[3] J.R. Deller, J.G.Proakis, J.H.L.Hansen, "Discrete-Time Processing of Speech Signals", Macmillan 1993.
[4] R. Mammone, X. Zhang, R. Ramachandran, "Robust Speaker Recognition – a feature based approach", IEEE Signal Processing Magazine, vol 13, no 5, 1996.
[5] S. Furui, "Cepstral Analysis technique for automatic speaker verification", IEEE Transactions on Acoustics, Speech and Signal Processing, vol. ASSP-29, 1981.
[6] Mike Brooks, VOICEBOX: Speech Processing Toolbox for MATLAB, Imperial College, available online: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html