ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΣΤΑΤΙΣΤΙΚΗ ΕΠΕΞΕΡΓΑΣΙΑ ΣΗΜΑΤΟΣ ΟΜΙΛΙΑΣ ΜΕ ΕΦΑΡΜΟΓΗ ΣΤΟ ΔΙΑΧΩΡΙΣΜΟ ΟΜΙΛΗΤΩΝ

Γιώργος Σιαντίκος

Φιλία Βόντα
Επίκ. Καθηγήτρια
ΣΕΜΦΕ, ΕΜΠ

Σέργιος Πετρίδης
Ερευνητής
ΕΚΕΦΕ Δημόκριτος

Αθήνα 2013

Περιεχόμενα

1 Επεξεργασία Ομιλίας **1**

1.1 Εισαγωγή 1

1.2 Ακοή & Παραγωγή Ομιλίας 3

1.2.1 Ακοή 3

1.2.2 Παραγωγή ομιλίας 9

1.3 Βραχύχρονη Ανάλυση & Εξαγωγή Χαρακτηριστικών 13

1.3.1 Απλά παραδείγματα 15

1.4 Μετασχηματισμός Fourier Βραχέος Χρόνου 20

1.4.1 Σειρές Fourier 20

1.4.2 Μετασχηματισμός Fourier 22

1.4.3 Διακριτός Μετασχηματισμός Fourier 23

1.5 Γραμμική Πρόβλεψη 34

1.5.1 Μοντέλο ιδανικών σωλήνων 34

1.5.2 Εκτίμηση Παραμέτρων 35

1.5.3 Ισοδύναμες μορφές 40

1.6 Ομομορφική Ανάλυση 44

2 Ομαδοποίηση δεδομένων & Μείωση Διαστάσεων **51**

2.1 Ομαδοποίηση Δεδομένων 51

2.1.1 Η μέθοδος K-μέσων 54

2.1.2 Μέθοδος Ασαφών K-μέσων 57

2.1.3 Επαλήθευση Ομαδοποίησης 63

2.2 Μείωση Διαστάσεων 65

2.2.1 Ανάλυση Κύριων Συνιστωσών 66
2.2.2 Γραμμική Διαχωριστική Ανάλυση 72

3 Διαχωρισμός Ομιλητών 79
3.1 Τυπικές Μέθοδοι Διαχωρισμού Ομιλητών 81
3.1.1 Ανίχνευση ενός σημείου αλλαγής 82
3.1.2 Ανίχνευση πολλών σημείων αλλαγής 85
3.2 Η μέθοδος FLsD 90
3.2.1 Θεωρία 91
3.2.2 Εφαρμογή 95
3.2.3 diarx 102

Βιβλιογραφία 109

Περίληψη

Σε αυτή την εργασία περιγράφονται βασικές μέθοδοι επεξεργασίας ομιλίας, καθώς και τεχνικές που χρησιμοποιούνται σε μία από τις κύριες εφαρμογές της, το *Διαχωρισμό Ομιλητών*, όπου καλούμαστε να απαντήσουμε στο ερώτημα *"ποιος μίλησε πότε;"*. Παρουσιάζονται τεχνικές που έχουν χρησιμοποιηθεί στο παρελθόν, ενώ σημείο αναφοράς είναι μία από τις πιο πρόσφατες μεθόδους που αναπτύχθηκε τα τελευταία χρόνια στο ΕΚΕΦΕ Δημόκριτος. Παράλληλα με τη συγγραφή του κειμένου, αναπτύχθηκε και το σχετικό λογισμικό **diarx** σε γλώσσα C++, με το οποίο έγιναν δοκιμές σε αρχεία της βιβλιοθήκης του Canal9. Τμήματά του μπορούν να βρεθούν στην ηλεκτρονική διεύθυνση (αποθετήριο git)

https://gntikos@bitbucket.org/gntikos/diarx.git

Abstract

We describe some basic techniques for speech processing and one of its main applications, *Speaker Diarization*, where we try to answer the question *"who spoke when?"*. We present some techniques traditionally used for this task, while we focus on a state-of-the-art diarization method developed in IIT Demokritos during the last years. We also developed the relevant software **diarx** in C++ language, which was used to test the method in real-world speech content taken from the Canal9 corpus. Parts of the software are available through its git repository

https://gntikos@bitbucket.org/gntikos/diarx.git

ΕΥΧΑΡΙΣΤΙΕΣ

Για αυτή την εργασία, θέλω να ευχαριστήσω τον κύριο Σέργιο Πετρίδη, κυρίως για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα από τα πιο ενδιαφέροντα κατά τη γνώμη μου θέματα, αλλά και για την εμπιστοσύνη και υπομονή που έδειξε κατά τη διάρκεια εκπόνησής της. Επίσης, ευχαριστώ την κυρία Φιλία Βόντα που με διευκόλυνε στο μέγιστο βαθμό στην ολοκλήρωση της εργασίας με την προθυμία της να αναλάβει το ρόλο της επιβλέπουσας καθηγήτριας, τις διορθώσεις και την καλή διάθεση με την οποία αντιμετώπισε τις δυσκολίες που προέκυψαν . Τέλος ευχαριστώ τους κυρίους Κουρκουλή και Σπαθή και τη γραμματεία ΣΕΜΦΕ, που μου προσέφεραν 3 χρόνια αδράνειας έως την αποφοίτησή μου.

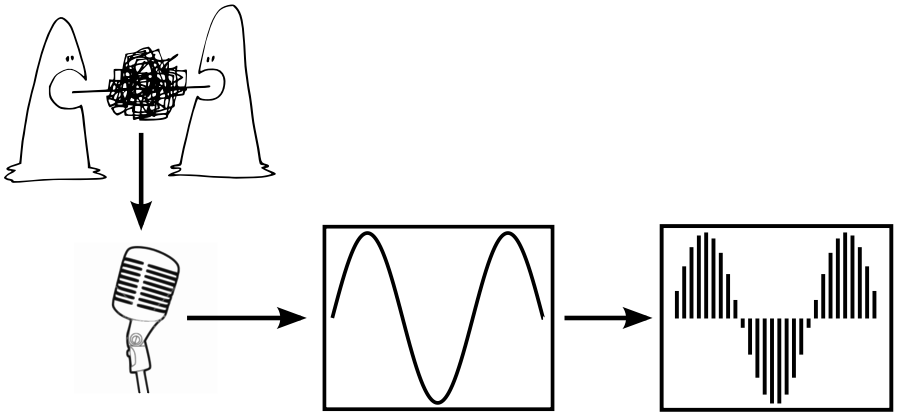
ΚΕΦΑΛΑΙΟ 1

Επεξεργασία Ομιλίας

1.1 Εισαγωγή

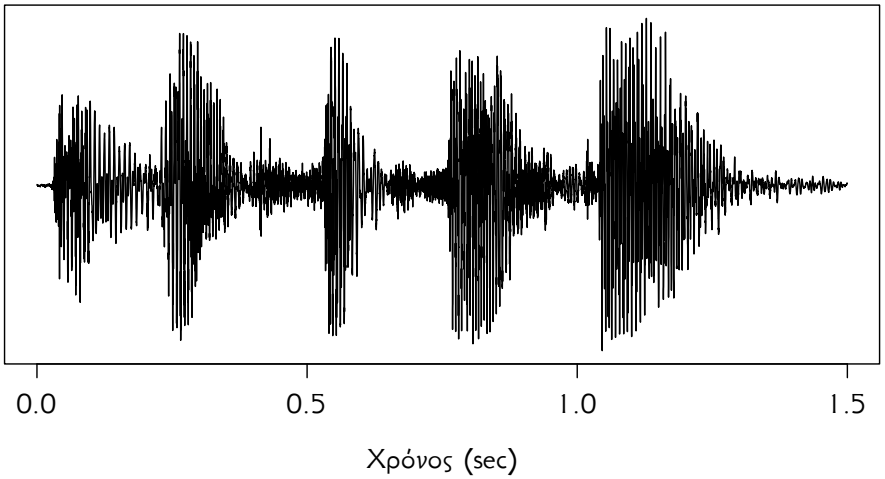
Η ομιλία είναι ο βασικός τρόπος επικοινωνίας των ανθρώπων. Συνεπώς η ικανότητα επεξεργασίας της, με την έννοια της άντλησης πληροφορίας και νοήματος από αυτή, είναι ζωτικής σημασίας για τους ανθρώπους και τις κοινωνίες. Ενώ όμως στην καθημερινότητα αυτή η επεξεργασία γίνεται με μηχανικό τρόπο από τον καθένα στις συνομιλίες που έχει ή στις ομιλίες που παρακολουθεί μέσω της ακρόασης, ο τεράστιος διαθέσιμος όγκος *καταγεγραμμένης* ομιλίας ωθεί στην ανάπτυξη αυτόματων μεθόδων επεξεργασίας καθώς ο χρόνος και ο κόπος που απαιτεί η ακρόαση είναι απαγορευτικά μεγάλοι. Ενώ όμως τα μέσα για καταγραφή ομιλίας και γενικότερα ήχου υπήρχαν ήδη από τα μέσα του 19^{ου} αιώνα [Milner, 2009], η ανάγκη αυτόματων μεθόδων δημιουργήθηκε πολύ αργότερα με την ανάπτυξη συστημάτων *ψηφιακής* καταγραφής κατά την οποία το μηχανικό κύμα αέρα που αποτελεί την ομιλία, μετατρέπεται σε ηλεκτρικό μέσω ενός μικροφώνου και στη συνέχεια μέσω δειγματοληψίας σε μία ακολουθία αριθμών η οποία μπορεί να αποθηκευθεί σε αρχεία για μελλοντική χρήση (Σχήμα 1.1). Με την πάροδο του χρόνου αυτή η μετατροπή έχει γίνει εύκολη και οικονομική και το ίδιο ισχύει για την αποθήκευση αλλά και τη διάθεση των αρχείων δημιουργώντας έτσι το μεγάλο όγκο δεδομένων που αναφέραμε.

Η *ψηφιακή επεξεργασία ομιλίας* είναι ένα σχετικά νέο επιστημονικό πεδίο με αμέ-



Σχήμα 1.1: Ψηφιοποίηση ομιλίας

τρητες εφαρμογές που διαρκώς αυξάνονται καθώς το ερευνητικό ενδιαφέρον παραμένει υψηλό. Το προς επεξεργασία αντικείμενο εδώ δεν είναι η ίδια η ομιλία, αλλά η ακολουθία αριθμών που προέρχεται από την ψηφιοποίησή της (Σχήμα 1.2) και έτσι οι τεχνικές που χρησιμοποιούνται συναντώνται και σε πολλές εφαρμογές της Ανάλυσης Χρονοσειρών ή της Ανάλυσης Σήματος, με μικρές όμως τροποποιήσεις ώστε να λαμβάνονται υπόψη οι ιδιαιτερότητες ενός σήματος ομιλίας.



Σχήμα 1.2: Απόσπασμα ομιλίας Α. Σαμαρά /a Greek success story/ (πηγή: www.enikos.gr)

1.2 Ακοή & Παραγωγή Ομιλίας

Πριν δείξουμε μερικές από τις βασικές τεχνικές που χρησιμοποιούνται, είναι απαραίτητες κάποιες πληροφορίες για τον ανθρώπινο οργανισμό, συγκεκριμένα για το μηχανισμό ακοής και το μηχανισμό παραγωγής ομιλίας: Η αντίληψη του ήχου είναι ένα πολύ περίπλοκο θέμα και ακόμα και σήμερα δεν έχει διασαφηνιστεί πλήρως. Σε κάθε περίπτωση όμως, από τη στιγμή που στόχος μας είναι η δημιουργία μεθόδων που κάνουν αυτό που κάνει ο ανθρώπινος οργανισμός, κάποια γνώση για τον τρόπο λειτουργίας του σίγουρα είναι χρήσιμη. Όσο για το μηχανισμό παραγωγής ομιλίας, η ανάγκη γνώσης της λειτουργίας του είναι προφανής αφού μέσω αυτής μπορούμε να καταλάβουμε τη μορφή του σήματος ομιλίας που διαχειριζόμαστε. Πέρα από αυτό όμως, γνωρίζοντας το πως δουλεύει ο μηχανισμός ομιλίας μπορούμε να κατασκευάσουμε κατάλληλα μοντέλα τα οποία εκμεταλλεύονται τις ιδιαιτερότητές του και βοηθούν στην επεξεργασία.

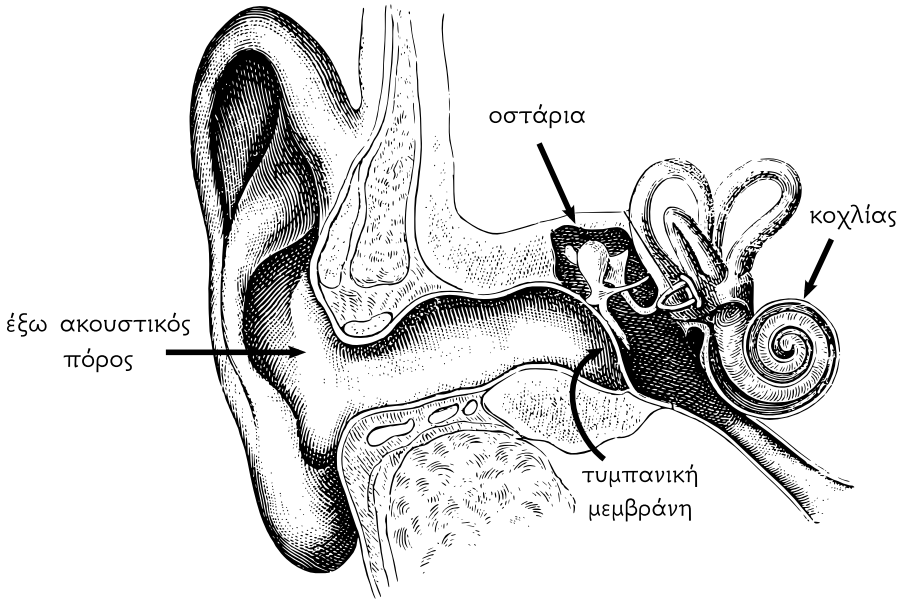
1.2.1 Ακοή

Ο ήχος είναι ένα διαμήκες κύμα που διαδίδεται στον αέρα και ως τέτοιο χαρακτηρίζεται κάθε στιγμή από το πλάτος και τη συχνότητά του. Αυτό που διακρίνει τον ήχο από ένα άλλο διαμήκες κύμα είναι ότι τα χαρακτηριστικά του είναι κατάλληλα ώστε να γίνει αντιληπτός από το ανθρώπινο αυτί που είναι το αισθητήριο όργανο ακοής (Σχήμα 1.3). Η λειτουργία της ακοής είναι περίπλοκη καθώς το ίδιο το αυτί είναι ένα πολύ σύνθετο όργανο. Συνοπτικά πάντως έχει ως εξής: Ο αέρας εισέρχεται στο αυτί μέσω του *εξωτερικού σωλήνα* και πιέζει την *τυμπανική μεμβράνη*. Αυτή με τη σειρά της κινεί μία σειρά μικρών οστών που εφάπτεται από τη μία μεριά στο εσωτερικό της τυμπανικής μεμβράνης και από την άλλη στο *οβάλ παράθυρο* του *κοχλία*. Ο κοχλίας (Σχήμα 1.4) είναι και το σημαντικότερο τμήμα του αυτιού όσον αφορά την ακοή. Το εσωτερικό του είναι χωρισμένο σε δύο θαλάμους που περιέχουν υγρό και μεταξύ τους βρίσκεται η *βασική μεμβράνη*. Η διαταραχή που προκαλείται από τα οστάρια στο υγρό, ταξιδεύει προς την *κορυφή* (και μετά προς τα πίσω μέσω του άλλου θαλάμου), προκαλώντας ταλαντώσεις στη βασική μεμβράνη. Οι διαστάσεις της βασικής μεμβράνης δεν είναι σταθερές σε όλη αυτή τη διαδρομή και έτσι διαφορετικά

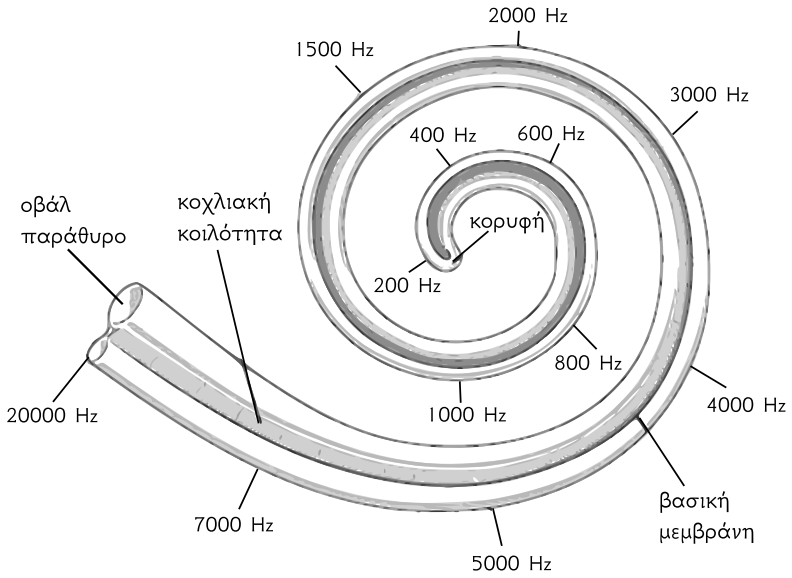
μέρη της διεγείρονται από διαφορετικές συχνότητες. Από τη διέγερση αυτή δημιουργείται ένας ηλεκτρικός παλμός ο οποίος κατευθύνεται προς τον εγκέφαλο όπου γίνεται η "μετάφραση" του ηλεκτρικού σήματος σε ηχητικό μήνυμα.

Η φυσιολογία του αυτιού και των εσωτερικών οργάνων του είναι αυτές που προσδιορίζουν τις ιδιότητες που πρέπει να έχει ένα κύμα για να γίνεται αντιληπτό ως ήχος. Δεδομένων των μικροδιαφορών που υπάρχουν μεταξύ ατόμων οι ιδιότητες αυτές δεν είναι ίδιες για τον καθένα αλλά σε γενικές γραμμές, οι τιμές που παρουσιάζουμε αν και ενδεικτικές, μπορούν να γίνουν αποδεκτές. Οι συχνότητες που είναι ικανές να διεγείρουν τη βασική μεμβράνη του κοχλία ορίζουν το φάσμα ακουστότητας ως το διάστημα 20 Hz - 20 kHz.

Αυτή είναι μία πολύ σημαντική παρατήρηση που έχει επιπτώσεις στην ψηφιοποίηση του ήχου. Μπορούμε να φανταστούμε ότι σε μία χρονοσειρά όπου ο χρόνος είναι διακριτός, δεν μπορούν να αναπαρασταθούν σωστά όλες οι συχνότητες. Συνεπώς, η δειγματοληψία πρέπει να γίνει με ένα ρυθμό f_s που να επιτρέπει την παρουσία όλων των επιθυμητών συχνοτήτων. Το παρακάτω θεώρημα εξηγεί αυτούς τους περιορισμούς καλύτερα.



Σχήμα 1.3: Το ανθρώπινο αυτί (εικόνα από web.mit.edu)



Σχήμα 1.4: Ο κοχλίας του αυτιού (εικόνα από Encyclopaedia Britannica)

Θεώρημα 1.2.1 (Θεώρημα Δειγματοληψίας Shannon-Nyquist [Shannon, 1949]). *Ένα σήμα $x_c(t)$ συνεχούς χρόνου το οποίο δεν περιέχει συχνότητες μεγαλύτερες της f_M μπορεί να προσδιοριστεί πλήρως από τα δείγματα $x[n] = x_c(nT_s)$, εάν ο ρυθμός δειγματοληψίας $f_s = 1/T_s$ ικανοποιεί τη σχέση $f_s \geq 2f_M$*

Απόδειξη. Έστω $X(e^{i\omega})$ ο μετασχηματισμός Fourier του σήματος $x_c(t)$. Τότε

$$\begin{aligned} x_c(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} X(e^{i\omega})e^{i\omega t}d\omega \\ &= \frac{1}{2\pi}\int_{-2\pi f_M}^{2\pi f_M} X(e^{i\omega})e^{i\omega t}\,d\omega \end{aligned} \tag{1.1}$$

αφού από υπόθεση $X(e^{i\omega}) \equiv 0$ για $|\omega| > 2\pi f_M$. Δειγματοληπτώντας σε διαστήματα μήκους

$$T_s = \frac{1}{2f_M} \tag{1.2}$$

ή ισοδύναμα, για χρόνους

$$t = \frac{n}{2f_M}, \quad n \in \mathbb{Z} \tag{1.3}$$

έχουμε

$$x[n] = x_c\left(\frac{n}{2f_M}\right) = \frac{1}{2\pi}\int_{-2\pi f_M}^{2\pi f_M} X(e^{i\omega}) \cdot \exp\left\{\frac{in\omega}{2f_M}\right\} d\omega \tag{1.4}$$

Παρατηρούμε ότι το ολοκλήρωμα στο δεξιό μέρος της τελευταίας εξίσωσης είναι ο n-οστός παράγοντας του αναπτύγματος Fourier της συνάρτησης $X(\omega)$ εάν θεωρήσουμε το διάστημα $(-f_M, f_M)$ αντιστοιχεί θεμελιώδη περίοδο. Έτσι το $X(\omega)$ προδιορίζεται πλήρως από τα δείγματα $x[n]$, αφού στα υπόλοιπα σημεία μηδενίζεται και συνεπώς προσδιορίζεται πλήρως και το αρχικό σήμα $x_c(t)$ μέσω του αντίστροφου μετασχηματισμού Fourier. □

Η ποσότητα $f_s/2$ ονομάζεται *συχνότητα Nyquist* και είναι η μέγιστη συχνότητα που μπορεί να αναπρασταθεί μέσω μετασχηματισμού Fourier όταν η συχνότητα δειγματοληψίας είναι f_s. Αυτό είναι και ένας οδηγός για τον τρόπο με τον οποίο ψηφιοποιείται ένα αναλογικό σήμα. Για παράδειγμα σε ένα CD μουσικής η συχνότητα δειγματοληψίας είναι 44100 Hz αφού η πιστότητα του ήχου έχει μεγάλη σημασία, ενώ σε αρχεία βίντεο για διαδικτυακή χρήση όπου το κόστος αποθήκευσης και η ταχύτητα ανάσυρσης παίζουν επίσης σημαντικό ρόλο χρησιμοποιούνται πολύ χαμηλότεροι ρυθμοί π.χ 22050 ή 16000 Hz. Όταν το αντικείμενο είναι η δειγματοληψία ανθρώπινης φωνής (π.χ σε εφαρμογές VOIP), το φασματικό εύρος μπορεί να περιοριστεί αρκετά αφού το εύρος των θεμελιωδών συχνοτήτων της ανθρώπινης φωνής είναι περίπου 80 Hz - 2000 Hz και έτσι σε τέτοιου είδους εφαρμογές συνήθως συναντάμε ρυθμούς δειγματοληψίας μεταξύ 4 kHz και 16 kHz.

Το άλλο σημαντικό μέγεθος που χαρακτηρίζει ένα κύμα ως ήχο είναι το πλάτος. Μετρώντας σε μονάδες πίεσης, οι τιμές του πλάτους κυμαίνονται από την ελάχιστη πίεση P_0 που πρέπει να ασκηθεί στην τυμπανική μεμβράνη για να τεθεί σε κίνηση έως τη μέγιστη που μπορεί να αντέξει χωρίς να τραυματιστεί. Επειδή το εύρος αυτό είναι πολύ μεγάλο, συνηθίζεται η μέτρησή του σε λογαριθμική κλίμακα και ειδικότερα σε *decibel (dB)*. Συγκεκριμένα, χρησιμοποιώντας την κλίμακα decibel ως τρόπο σύγκρισης δύο ήχων με επίπεδα ισχύος P_1 και P_2

$$10\log_{10}(P_1/P_2)$$

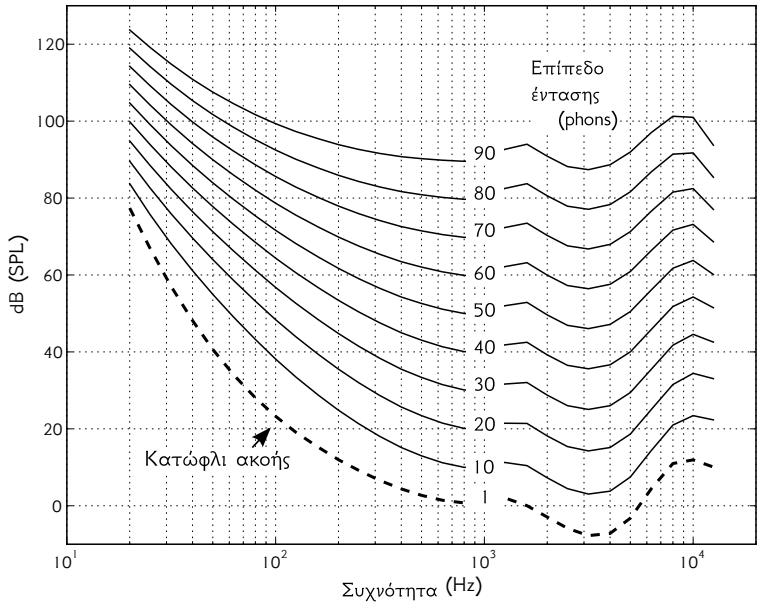
ορίζουμε το *Επίπεδο Ηχητικής Πίεσης (**S**ound **P**ressure **L**evel - SPL)*, ως ένα μέτρο απόλυτης ηχητικής πίεσης P σε dB:

$$SPL(dB) = 20\log_{10}\left(\frac{P}{P_0}\right) \tag{1.5}$$

όπου το σημείο αναφοράς 0 dB αντιστοιχεί στο κατώφλι της ακοής το οποίο είναι $P_0 = 2 \times 10^{-4}$ μbar για έναν ήχο συχνότητας 1 kHz. Ενδεικτικά, αναφέρουμε ότι το επίπεδο ομιλίας σε μία συνηθισμένη κουβέντα όπως ακούγεται από την απόσταση ενός μέτρου, είναι περίπου 60 dB SPL, ενώ το επίπεδο του ήχου ενός τρυπανιού περίπου 120 dB SPL.

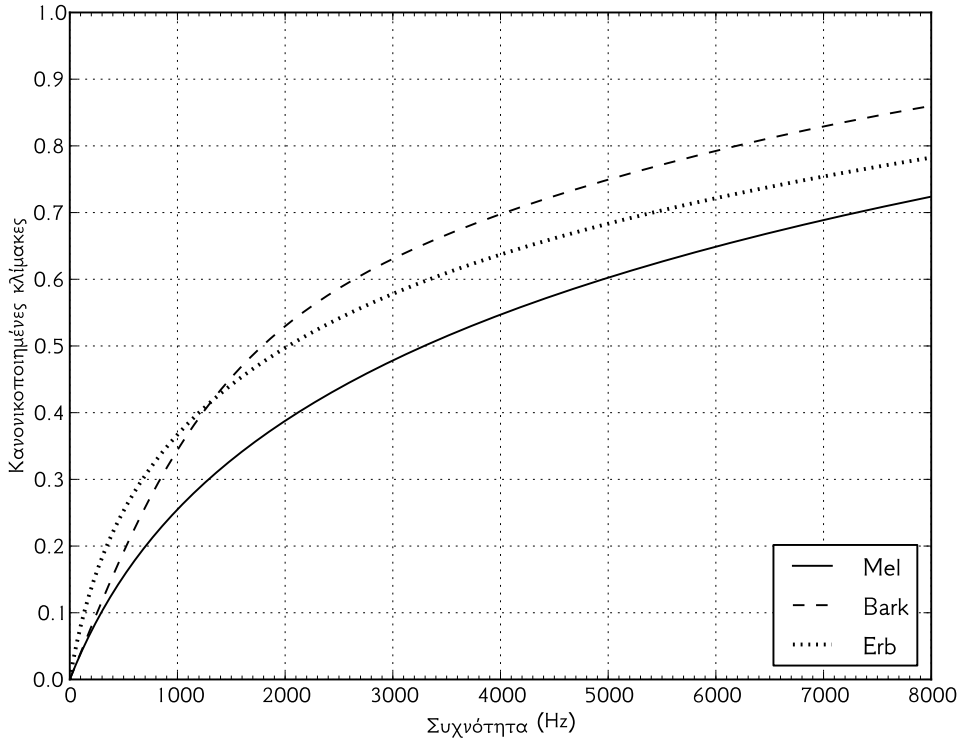
Έχει φανεί πειραματικά, ότι η απαιτούμενη ενέργεια για να γίνει αντιληπτός ένας ήχος είναι συνάρτηση της συχνότητάς του και αυτό οδηγεί σε μια διαφορετική μονάδα έντασης του ήχου, το *phon*. Αυτή η μονάδα δεν έχει σαν σκοπό τη μέτρηση των φυσικών ιδιοτήτων του ηχητικού κύματος, αλλά τη μέτρηση της έντασης του ήχου *όπως αυτή γίνεται αντιληπτή από έναν ακροατή* και έτσι 1 phon ορίζεται ως η ένταση ήχου που αισθάνεται ένας ακροατής όταν ακούει έναν τόνο 1 kHz σε επίπεδο 1 dB SPL.

Αυτή η μη-γραμμική ως προς τη συχνότητα απόκριση του αυτιού σε ήχους ίδιας έντασης, έχει σαν αποτέλεσμα τη συχνή υιοθέτηση εναλλακτικών, εμπειρικών



Σχήμα 1.5: Ισοϋψείς καμπύλες επιπέδων έντασης (βάσει του ISO 226)

κλιμάκων συχνότητας που αντανακλούν αυτή τη συμπεριφορά (Σχήμα 1.6), κάνοντας χρήση γνώσεων περί των *κρίσιμων ευρών ζώνης*, δηλαδή των περιοχών γύρω από μία κεντρική συχνότητα όπου όλες οι συχνότητες ηχούν το ίδιο. Όπως είπαμε, αυτές οι



Σχήμα 1.6: Εναλλακτικές κλίμακες

κλίμακες είναι εμπειρικές και κατασκευάζονται μέσω πειραματικής διαδικασίας. Φυσικά, από τη στιγμή που κάθε άτομο αντιλαμβάνεται διαφορετικά τον ήχο, οι κλίμακες είναι υποκειμενικές (όπως και η μονάδα phon), αλλά σε γενικές γραμμές μπορούμε να υποθέσουμε ότι προσεγγίζουν ικανοποιητικά τη μέση αντίληψη. Μία από αυτές είναι η *κλίμακα mel*, που κατασκευάστηκε από τους Stevens, Volkman και Newman το 1937 [Stevens et al., 1937]. Η γενική παρατήρηση που έγινε στα πειράματά τους είναι ότι η σχέση μεταξύ πραγματικής συχνότητας και αίσθησης είναι σχεδόν γραμμική για συχνότητες μικρότερες του 1 kHz και σχεδόν λογαριθμική μετά. Φυσικά δεν υπάρχει ακριβής τύπος για την κλίμακα mel [Umesh et al., 1999] αλλά μία από

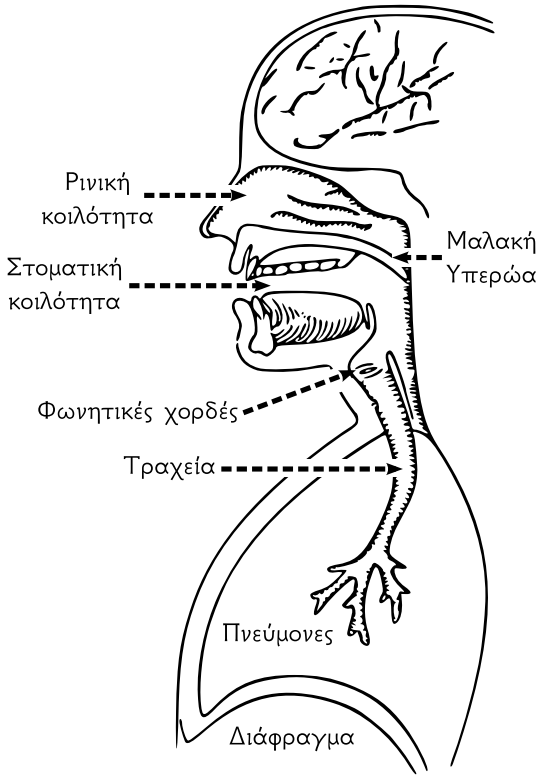
τις πιο δημοφιλείς προσεγγίσεις της δίνεται από τον τύπο

$$mel(f) = 2595 \times \log_{10}\left(\frac{f}{700} + 1\right) \tag{1.6}$$

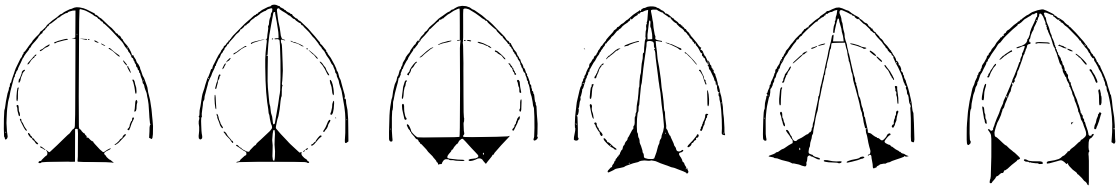
Βλέπουμε πως η λειτουργία του αυτιού είναι μία πολύ περίπλοκη διαδικασία, όπου εκτός από τους παράγοντες της φυσιολογίας των αισθητήριων οργάνων, εμπλέκονται και ψυχολογικοί παράγοντες, πράγμα που κάνει την ακριβή μελέτη του πρακτικά αδύνατη. Σε κάθε περίπτωση όμως, οι κλίμακες που παρουσιάσαμε είναι χρήσιμα και αποτελεσματικά εργαλεία για συστηματική χρήση.

1.2.2 Παραγωγή ομιλίας

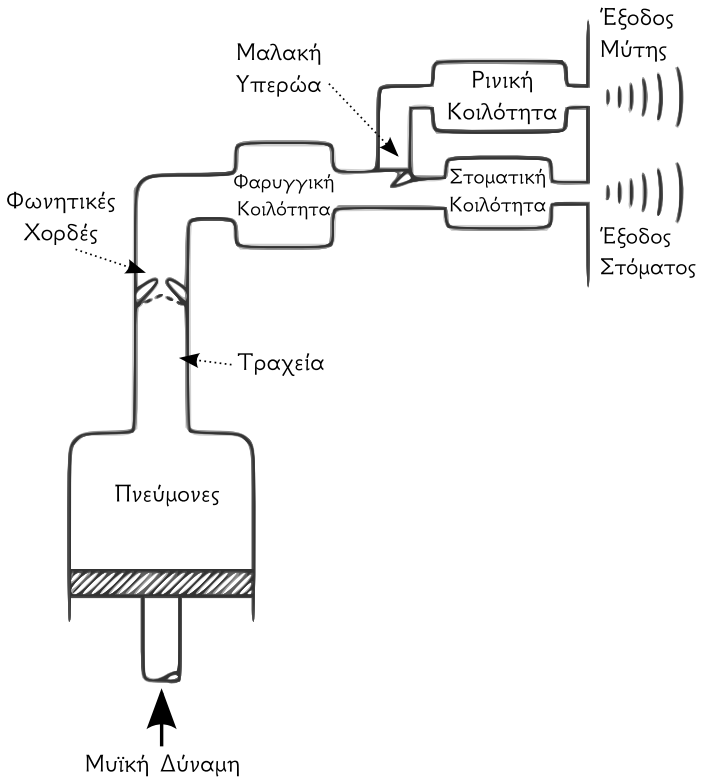
Προχωράμε στην περιγραφή του ανθρώπινου μηχανισμού παραγωγής ομιλίας (Σχήμα 1.7). Και εδώ, η ομιλία μελετάται ως ένα καθαρά μηχανικό φαινόμενο, απαλλαγμένο από κάθε είδους περιεχόμενο. Η ομιλία λοιπόν είναι απλά μία ροή αέρα που εξέρχεται από τη μύτη και το στόμα. Συνοπτικά, η διαδρομή του ρεύματος αέρα είναι η εξής: Η αφετηρία βρίσκεται στους *πνεύμονες*, όπου ο εκεί αποθηκευμένος αέρας οδηγείται προς την *τραχεία* πιεζόμενος από το *διάφραγμα*. Μέσα στην τραχεία ο αέρας δεν συναντά ιδιαίτερα εμπόδια και η ροή του μπορεί να θεωρηθεί στρωτή. Το πρώτο ενδιαφέρον σημείο της διαδρομής βρίσκεται στις *φωνητικές χορδές* (Σχήμα 1.8). Οι φωνητικές χορδές μπορούν να παρομοιαστούν με "πύλες" που ανοιγοκλείνουν είτε περιοδικά, μετατρέποντας το συνεχές ρεύμα αέρα σε μια σειρά περιοδικών παλμών, είτε ακανόνιστα, μετατρέποντας τη ροή από στρωτή σε τυρβώδη. Στη συνέχεια, ο αέρας περνάει από τη *φαρυγγική κοιλότητα* και κατευθύνεται στη *στοματική κοιλότητα* αλλά και στη *ρινική κοιλότητα* μέσω άλλης μιας "πύλης", της *μαλακής υπερώας*. Σε αυτές τις κοιλότητες γίνεται και το μεγαλύτερο μέρος της διαμόρφωσης της φωνής, καθώς η έξοδος από αυτές (κοινώς η φωνή), είναι αποτέλεσμα περίπλοκων διαδρομών του αέρα σε αυτές και διαρκών ανακλάσεων στα τοιχώματά τους. Ένα μηχανικό ανάλογο του ανθρώπινου μηχανισμού παραγωγής φωνής φαίνεται στο Σχήμα 1.9. Η μεγάλη ποικιλία ήχων που μπορούν να παραχθούν από τον άνθρωπο οφείλεται ακριβώς στις αμέτρητες παραμέτρους που διαμορφώνουν τις κοιλότητες μετά τις φωνητικές χορδές όπως η θέση της γλώσσας, η μεταβολή του όγκου της στοματικής κοιλότητας με κίνηση του σαγονιού, η θέση των χειλιών κτλ. Η ακριβής καταγραφή



Σχήμα 1.7: Μέρη του μηχανισμού παραγωγής ομιλίας (εικόνα από openi.nlm.nih.gov)



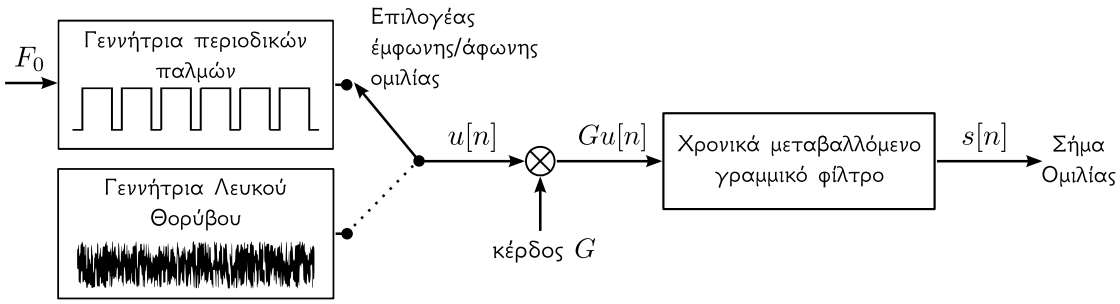
Σχήμα 1.8: Φάσεις κίνησης των φωνητικών χορδών (εικόνα από Wikimedia Commons)



Σχήμα 1.9: Μηχανικό ανάλογο του ανθρώπινου μηχανισμού παραγωγής φωνής [Flanagan, 1972]

και κατηγοριοποίηση των ήχων (ή φωνημάτων) που παράγονται είναι μεταξύ των αντικειμένων της γλωσσολογίας, αλλά όσον αφορά τη στατιστική επεξεργασία ομιλίας μπορούμε να χρησιμοποιήσουμε δύο ευρείες κατηγορίες: την *έμφωνη* και την *άφωνη* ομιλία. Η έμφωνη ομιλία αντιστοιχεί στις περιπτώσεις άρθρωσης όπου οι φωνητικές χορδές πάλλονται περιοδικά και περιλαμβάνει φθόγγους όπως τα φωνήεντα και τα ηχηρά σύμφωνα (π.χ. /z/), ενώ η άφωνη τις περιπτώσεις όπου οι φωνητικές χορδές πάλλονται ακανόνιστα, όπως κατά την άρθωση άηχων συμφώνων (π.χ /k/, /s/).

Όσον αφορά τη μοντελοποίηση της φωνητικής οδού, αν και δεν υπάρχουν οι δυσκολίες που υπάρχουν στη προσέγγιση των αισθητήριων οργάνων της ακοής, οι διαδρομές του αέρα είναι τόσο περίπλοκες που χρειάζεται ένα σαφώς απλουστευμένο μοντέλο. Ένα τέτοιο μοντέλο προτάθηκε από τον Gunnar Fant το 1960 [Fant, 1960] και ονομάζεται *μοντέλο πηγής-φίλτρου* (Σχήμα 1.10). Αυτό το μοντέλο έχει τύχει ευ-



Σχήμα 1.10: Μοντέλο μηχανισμού παραγωγής ομιλίας

ρείας αποδοχής από την ερευνητική κοινότητα, αφού παρ' όλη την απλότητά του, περιγράφει ικανοποιητικά τη λειτουργία της φωνητικής οδού. Για την παραγωγή ενός φθόγγου, γίνεται κατ' αρχήν μία επιλογή έμφωνης/άφωνης ομιλίας. Στην περίπτωση της έμφωνης ομιλίας, το ρεύμα αέρα από τις φωνητικές χορδές θεωρείται ως μία σειρά περιοδικών παλμών με θεμελιώδη συχνότητα F_0 που καθορίζει και το τονικό ύψος της φωνής, ενώ αντίθετα η τυρβώδης ροή αέρα κατά την άρθρωση άφωνης ομιλίας μοντελοποιείται ως χρονοσειρά λευκού θορύβου [1]. Η παράμετρος κέρδους G καθορίζει την ένταση της φωνής, ενώ το τελικό σήμα διαμορφώνεται με την εφαρμογή ενός γραμμικού φίλτρου που του οποίου η συμπεριφορά προσομοιάζει τις ανακλάσεις του αέρα στα τοιχώματα των διάφορων κοιλοτήτων της φωνητικής οδού. Η αποδοχή αυτού του μοντέλου ανάγει τη μοντελοποίηση της άρθρωσης ενός φθόγγου στην εκτίμηση των χαρακτηριστικών ενός γραμμικού φίλτρου, κάτι το οποίο μας απαλλάσσει από την ανάγκη προσδιορισμού των πολυάριθμων παραμέτρων της φωνητικής οδού ξεχωριστά. που άλλωστε δεν ενδιαφέρει (τουλάχιστον στη στατιστική ανάλυση σήματος ομιλίας).

[1] Καλούμε χρονοσειρά λευκού θορύβου μία στάσιμη χρονοσειρά ασυσχέτιστων τυχαίων μεταβλητών.

1.3 Βραχύχρονη Ανάλυση & Εξαγωγή Χαρακτηριστικών

Στην εκτίμηση των παραμέτρων ενός σήματος (ή γενικά μιας χρονοσειράς), κεντρικό ρόλο παίζει η έννοια της *στασιμότητας*. Με αυτόν τον όρο περιγράφουμε χρονοσειρές των οποίων οι βασικές στατιστικές ιδιότητες μένουν αναλλοίωτες στο χρόνο. Πιο συγκεκριμένα, ο ορισμός της στασιμότητας που συνήθως χρησιμοποιείται είναι ο εξής:

Ορισμός 1 (Ασθενής Στασιμότητα). *Μία χρονοσειρά* $\{X_t, t \in \mathbb{Z}\}$ *λέγεται* ασθενώς στάσιμη *εάν*

1. $\mu_t = \mathcal{E}\left[X_t\right] = \mu, \quad \forall t \in \mathbb{Z}$
2. $\gamma(t+h,t) = \gamma(h), \quad \forall t, h \in \mathbb{Z}$ *και*
3. $\sigma_t^2 = \mathrm{Var}\left[X_t\right] = \gamma(0) = \sigma^2 < \infty, \quad \forall t \in \mathbb{Z}$

αν δηλαδή οι συναρτήσεις μέσης τιμής και αυτοσυνδιακύμανσης είναι ανεξάρτητες του χρόνου t *και η συνάρτηση διασποράς πεπερασμένη.*

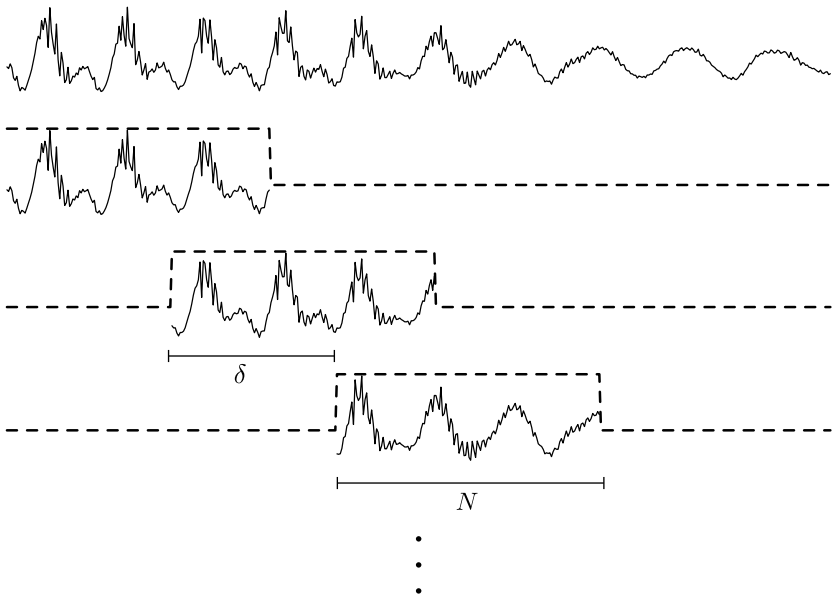
Φυσικά, το σήμα της ομιλίας δεν μπορεί να θεωρηθεί σε καμία περίπτωση στάσιμο (π.χ. Σχήμα 1.2), κάτι που μπορούμε βέβαια να αντιληφθούμε διαισθητικά, αλλά φαίνεται και από τον προσδιορισμό του γραμμικού φίλτρου του μοντέλου του Σχήματος 1.10 ως *χρονικά μεταβαλλόμενο*. Μία γενική παρατήρηση που μπορεί πάντως να γίνει, είναι ότι κατά τη διάρκεια άρθρωσης ενός φθόγγου, οι παράμετροι του μοντέλου πηγής-φίλτρου παραμένουν αμετάβλητες. Επίσης, ο ρυθμός μεταβολής του σήματος είναι κατά πολύ μεγαλύτερος του ρυθμού άρθρωσης. Για παράδειγμα, χρειάζονται 8000 δείγματα για να αναπαραστήσουν 1 δευτερόλεπτο ομιλίας με ρυθμό δειγματοληψίας 8000 Hz, ενώ σε αυτό το διάστημα σπάνια περιέχονται πάνω από 10 φθόγγοι. Βέβαια, όλοι οι φθόγγοι δεν έχουν την ίδια διάρκεια, αλλά είναι ασφαλές να δεχθούμε ότι για χρονικές διάρκειες περίπου 10 ms-30 ms το σήμα ομιλίας μπορεί να θεωρηθεί στάσιμο. Αυτή είναι και η βασική ιδέα της *βραχύχρονης ανάλυσης* ομιλίας, ότι η όποια επεξεργασία επιθυμούμε να εκτελέσουμε στο σήμα ομιλίας, γίνεται για διαδοχικά "μικρά" σε διάρκεια τμήματά της. Ο πιο απλός τρόπος για να εκφραστεί αυτό μαθηματικά, είναι ο πολλαπλασιασμός του σήματος με μία *συνάρτηση παραθύρου*

ή απλώς *παράθυρο* $w[n]$

$$w[n] = \begin{cases} 1 & 0 \leq x \leq N-1 \\ 0 & \text{αλλού} \end{cases} \tag{1.7}$$

που τοποθετείται κατάλληλα ώστε μόνο το επιθυμητό μήκους N τμήμα του σήματος να μην μηδενίζεται και στη συνέχεια μετακινείται κατά δ ώστε να αποκτηθεί το επόμενο τμήμα (Σχήμα 1.11). Το καθένα από αυτά τα τμήματα ομιλίας ονομάζεται *πλαίσιο ομιλίας*.

Ανάλογα με το είδος πληροφορίας που θέλουμε να πάρουμε από το σήμα ομιλίας, γίνεται και η αντίστοιχη επεξεργασία σε κάθε πλαίσιο ομιλίας για την εξαγωγή ενός κατάλληλου *χαρακτηριστικού* το οποίο συνήθως έχει τη μορφή ενός αριθμού ή ενός διανύσματος. Όλη η ουσία της επεξεργασίας ομιλίας είναι ακριβώς ο προσδιορισμός του κατάλληλου χαρακτηριστικού ή χαρακτηριστικών για την κάθε εργασία υλοποιώντας έτσι μια παραμετροποίηση του σήματος ομιλίας, καθώς και η επεξεργασία τους για την εξαγωγή συμπερασμάτων. Θα δείξουμε στη συνέχεια κάποιες βασικές τεχνικές επεξεργασίας σήματος ομιλίας και τονίζουμε ότι αυτές οι τεχνικές δεν εφαρμόζονται απ' ευθείας στο σήμα $x[n]$, αλλά σε διαδοχικά πλαίσιά του και για αυτό αρκετές φορές



Σχήμα 1.11: Διαδοχικά πλαίσια ομιλίας

τα ονόματα των χαρακτηριστικών περιέχουν τον προσδιορισμό "βραχέος χρόνου".

1.3.1 Απλά παραδείγματα

Δείχνουμε κάποια απλά παραδείγματα επεξεργασίας ομιλίας σύμφωνα με τις αρχές της βραχύχρονης ανάλυσης. Τα χαρακτηριστικά που θα δείξουμε μπορούν να εκφραστούν μαθηματικά με τη μορφή

$$Q_m = \sum_{n=-\infty}^{\infty} T\{x[n]\}w[n-m] \tag{1.8}$$

Το σήμα ομιλίας υπόκειται σε ένα μετασχηματιμό $\mathrm{T}\{\cdot\}$, ο οποίος δρα σε κάθε δείγμα χωριστά και και στη συνέχεια πολλαπλασιάζεται με ένα παράθυρο που είναι κατάλληλα τοποθετημένο κοντά στη θέση m. Τελικά αθροίζονται οι τιμές αυτού του γινομένου και έτσι επιτυγχάνεται η παραμετροποίηση των πλαισίων σαν μία χρονοσειρά $\{Q_m\}$. Το άπειρο άθροισμα υπάρχει για λόγους γενίκευσης και δεν χρειάζεται κάποια ιδιαίτερη προσοχή, αφού μπορεί στην ουσία είναι πεπερασμένο από τη στιγμή που είναι πεπερασμένος ο αριθμός των μη μηδενικών τιμών και μπορεί να γραφεί σαν

$$Q_m = \sum_{n=m-N+1}^{m} T\{x[n]\}w[n-m] \tag{1.9}$$

Ενέργεια Βραχέος Χρόνου

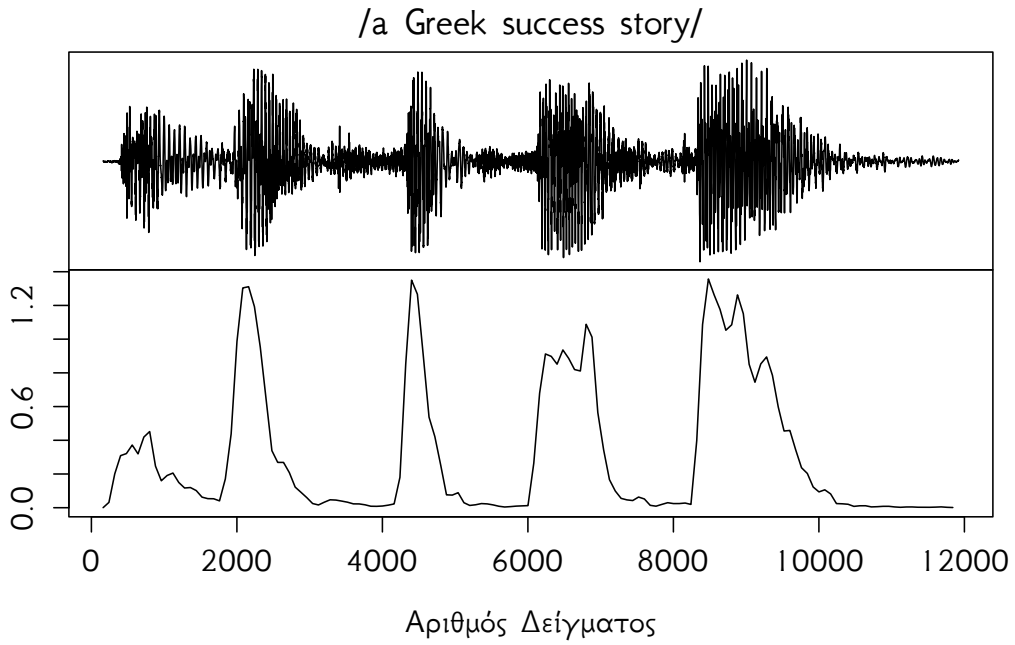
Η *ενέργεια βραχέος χρόνου* είναι ένα από τα πιο απλά παραδείγματα εφαρμογής της βραχύχρονης ανάλυσης ομιλίας. Ο τελεστής T ορίζεται από τη σχέση

$$T(x) = x^2 \tag{1.10}$$

και έτσι η ενέργεια δίνεται από την εξίσωση

$$E_m = \sum_{n=-\infty}^{\infty} (x[n]w[n-m])^2 \tag{1.11}$$

Μπορούμε να δούμε σχηματικά το αποτέλεσμα αυτού του μετασχηματισμού στο Σχήμα 1.12. Όσο απλό και αν είναι αυτό το χαρακτηριστικό, η πληροφορία που μας δίνει είναι πολύτιμη και για αυτό βρίσκει πολλές εφαρμογές: Μία υψηλή τιμή της



Σχήμα 1.12: /a Greek success story/: Ενέργεια βραχέος χρόνου

ενέργειας είναι ένδειξη ότι υπάρχει κάποια δραστηριότητα. Το περιεχόμενο αυτής της δραστηριότητας δεν μπορεί να εξαχθεί πλέον, αλλά σε κάθε περίπτωση μπορούμε να χωρίσουμε το σήμα σε περιοχές που "κάτι συμβαίνει" και σε περιοχές που πιθανόν μας είναι άχρηστες. Μία δημοφιλής εφαρμογή αυτού του διαχωρισμού είναι η διακοπή μετάδοσης σήματος (π.χ σε κινητά τηλέφωνα ή εφαρμογές VOIP) με σκοπό την εξοικονόμηση ενέργειας και εύρους ζώνης και υπολογιστικής ισχύος.

Μέσος Ρυθμός Διέλευσης από το μηδέν

Ένα ακόμη απλό χαρακτηριστικό του σήματος ομιλίας είναι ο *μέσος ρυθμός διέλευσης από το μηδέν* **A**verage (**Z**ero **C**rossing **R**ate - AZCR). Σε ένα σήμα διακριτού χρόνου, διέλευση από το μηδέν πρακτικά σημαίνει αλλαγή προσήμου, έτσι ένας κατάλληλος ορισμός δίνεται από την εξίσωση

$$Z_m = \frac{1}{N} \sum_{n=-\infty}^{\infty} |sgn(x[m]) - sgn(x[m-1])| \, w[n-m] \tag{1.12}$$

όπου

$$sgn(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases} \tag{1.13}$$

Ο ρυθμός αλλαγής διέλευσης από το μηδέν είναι ένα απλό μέτρο του συχνοτικού περιεχομένου ενός σήματος, ιδιαίτερα όταν αυτό είναι περιορισμένο. Για παράδειγμα, για ένα ημιτονοειδές σήμα συχνότητας f_0, όταν γίνει δειγματοληψία με ρυθμό f_s, αντιστοιχούν f_s/f_0 δείγματα σε κάθε κύκλο του. Σε κάθε κύκλο υπάρχουν δύο διελεύσεις από το μηδέν, άρα ο μέσος όρος διελεύσεων θα είναι

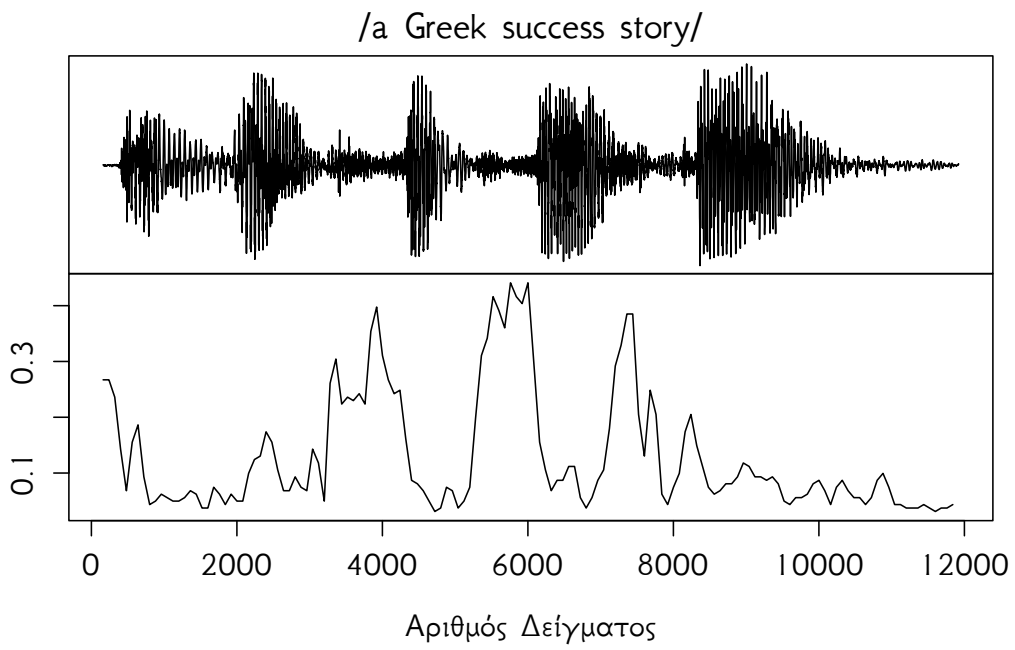
$$Z = 2\, f_0 / f_s \ \text{ διελεύσεις/δείγμα} \tag{1.14}$$

Το συχνοτικό περιεχόμενο των σημάτων ομιλίας είναι φυσικά πολύ πιο ευρύ και η εκτίμησή του από ένα τέτοιο απλό μέτρο είναι ακατάλληλη. Παρ' όλα αυτά μπορούμε να αντλήσουμε κάποιες πληροφορίες, κάνοντας την παρατήρηση ότι οι υψηλές συχνότητες συνεπάγονται υψηλούς ρυθμούς διέλευσης απ' το μηδέν ενώ οι χαμηλές το αντίθετο και άρα περιμένουμε να υπάρχει κάποιου είδους σχέση μεταξύ των μέσων ρυθμών διέλευσης και της κατανομής ενέργειας στις συχνότητες. Μία γενίκευση που μπορεί να γίνει, είναι ότι οι υψηλές τιμές του μέσου ρυθμού διέλευσης αντιστοιχεί σε τμήματα άφωνης ομιλίας, και οι χαμηλές σε έμφωνη. Η τελευταία πρόταση δεν είναι συγκεκριμένη, αλλά από το Σχήμα 1.13 εύκολα φαίνονται τα τμήματα έμφωνης και άφωνης ομιλίας.

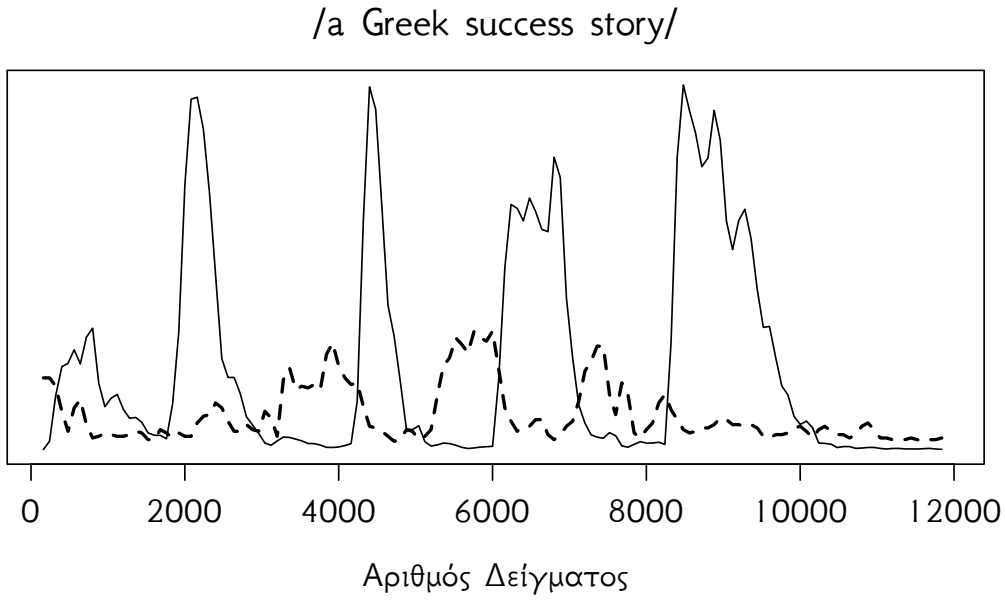
Επιπλέον παρατηρήσεις

Η ενέργεια βραχέος χρόνου μπορεί επίσης να λειτουργήσει (αν και σε μικρότερο βαθμό) σαν δείκτης έμφωνης/άφωνης ομιλίας αφού στις περιοχές έμφωνης ομιλίας η τιμή της είναι μεγαλύτερη, κάτι που φαίνεται και στο Σχήμα 1.14 όπου αντιπαρατίθεται με το μέσο ρυθμό διέλευσης από το μηδέν. Μάλιστα υπάρχουν αλγόριθμοι διαχωρισμού έμφωνης/άφωνης ομιλίας βασισμένοι σε αυτή την παρατήρηση, π.χ. [Bachu et al., 2010].

Κάτι που δεν γίνεται άμεσα εμφανές από τα σχήματα και ισχύει για κάθε χαρακτηριστικό που λαμβάνεται με παρόμοιο τρόπο, είναι ότι η μετασχηματισμένη χρονοσειρά

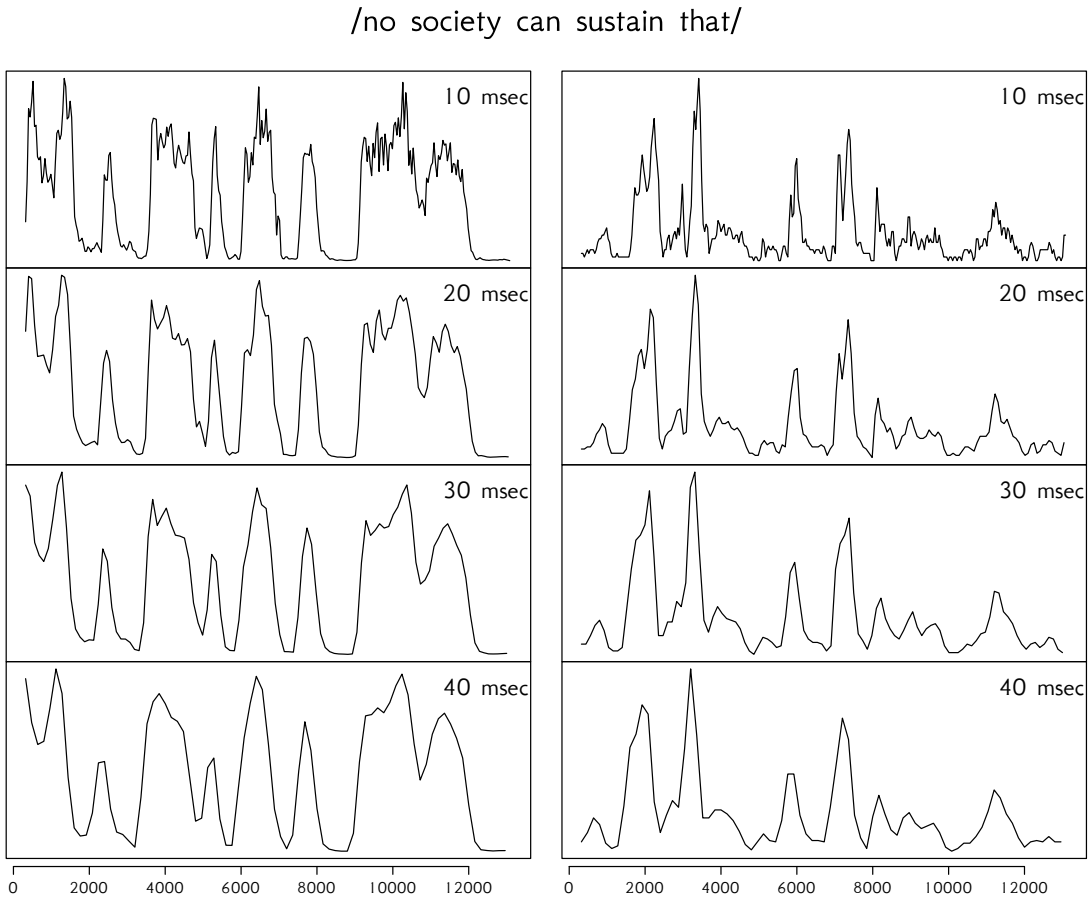


Σχήμα 1.13: /a Greek success story/: Μέσος ρυθμός διέλευσης από το μηδέν



Σχήμα 1.14: Αντιπαράθεση των δύο χαρακτηριστικών: η συνεχής γραμμή αντιστοιχεί στην ενέργεια βραχέος χρόνου και η διακεκομμένη στο μέσο ρυθμό διέλευσης απ' το μηδέν.

είναι κατά τάξεις μεγέθους μικρότερη από την αρχική και κατά συνέπεια η επεξεργασία της, αποθήκευσή της κτλ. πολύ πιο "οικονομική". Είπαμε πως "λογικές" τιμές για μήκη παραθύρου είναι μεταξυ 10 και 30 msec επειδή σε τέτοια διαστήματα μπορούμε να θεωρήσουμε σταθερές τις παραμέτρους της φωνητικής οδού. Κάτι τέτοιο όμως δεν χρειάζεται να ισχύει για τα χαρακτηριστικά που δείξαμε, έτσι το μήκος παραθύρου μπορεί να επιλεγεί σύμφωνα με τις ανάγκες μας σε χρονική ανάλυση. Στο Σχήμα 1.15 φαίνεται το αποτέλεσμα της επιλογής μικρότερων ή μεγαλύτερων παραθύρων.



Σχήμα 1.15: Επίδραση του μεγέθους παραθύρου στα χαρακτηριστικά (αριστερά η ενέργεια βραχέος χρόνου, δεξιά ο μέσος ρυθμός διέλευσης απ' το μηδέν).

Παρ' όλο που τα χαρακτηριστικά που παρουσιάσαμε είναι πολύ απλά, βρίσκουν εφαρμογές είτε μόνα τους είτε σε συνδυασμό με άλλα. Στη συνέχεια θα παρουσιάσουμε βασικές τεχνικές επεξεργασίας σήματος ομιλίας που οδηγούν σε κάπως πιο σύνθετα

χαρακτηριστικά. Αυτό που είναι σημαντικό να έχουμε κατά νου όμως, είναι ότι σε κάθε περίπτωση, η βασική ιδέα της βραχύχρονης ανάλυσης είναι παρούσα και η μόνη διαφορά μεταξύ των χαρακτηριστικών είναι η συνάρτηση που εφαρμόζεται στα πλαίσια ομιλίας.

1.4 Μετασχηματισμός Fourier Βραχέος Χρόνου

Οι *αναπαραστάσεις Fourier* έχουν παίξει παραδοσιακά σημαντικό ρόλο στο πεδίο της επεξεργασίας σήματος ομιλίας. Ο βασικός λόγος για αυτό έρχεται από το μοντέλο παραγωγής φωνής, όπου το σήμα διαμορφώνεται με την εφαρμογή ενός γραμμικού φίλτρου σε μία περιοδική ή τυχαία διέγερση. Είναι αναμενόμενο λοιπόν, τα χαρακτηριστικά και των δύο να είναι παρόντα στην αναπαράσταση Fourier η πιο απλά το *φάσμα* του σήματος.

Ο κεντρικός πυλώνας της θεωρίας Fourier είναι οι *σειρές Fourier* και αφού τις παρουσιάσουμε συνοπτικά μαζί με κάποιες γενικεύσεις τους, θα δούμε με ποιον τρόπο μπορούν αυτές οι τεχνικές να εφαρμοστούν στο σήμα ομιλίας.

1.4.1 Σειρές Fourier

Το 1807, ο Joseph Fourier εξέδωσε μια εργασία όπου παρουσίαζε τα αποτελέσματα της έρευνάς του για την επίλυση του προβλήματος Μερικών Διαφορικών Εξισώσεων που αφορούσε τη κατανομή της θερμοκρασίας. Σε αυτή την εργασία απέδειξε ότι κάθε συνάρτηση $x(t)$ που ικανοποιεί τις συνθήκες Dirichlet, δηλαδή

1. Η $x(t)$ είναι φραγμένη σε κάθε κλειστό διάστημα $[a, b]$ του $(-\infty, \infty)$

2. Η $x(t)$ έχει πεπερασμένο αριθμό μεγίστων και ελαχίστων σε κάθε διάστημα $[a, b]$

3. Η $x(t)$ είναι κατά διαστήματα συνεχής σε κάθε κλειστό διάστημα $[a, b]$.

4. Η $x(t)$ είναι περιοδική με περίοδο T, δηλαδή $x(t) = f(x + T)$ για κάθε $t \in \mathbb{R}$.

μπορεί να εκφραστεί ως άπειρο άθροισμα τριγωνομετρικών συναρτήσεων σύμφωνα με τη σχέση

$$x(t) = \frac{A_0}{2} + \sum_{k=1}^{\infty} \left[A_k \cos\left(\frac{2\pi kt}{T}\right) + B_k \sin\left(\frac{2\pi kt}{T}\right) \right] \tag{1.15}$$

όπου

$$A_k = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \cos\left(\frac{2\pi kt}{T}\right) dt, \quad k \geq 0 \tag{1.16α'}$$

$$B_k = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \sin\left(\frac{2\pi kt}{T}\right) dt, \quad k \geq 1 \tag{1.16β'}$$

Η σειρά της εξίσωσης (1.15) ονομάζεται *σειρά Fourier*. Σε κάθε σημείο ασυνέχειας t_a, η σειρά Fourier συγκλίνει στο μέσο όρο των δύο πλευρικών ορίων της $x(t)$, δηλαδή

$$\frac{x(t_a^+) + x(t_a^-)}{2} = \frac{A_0}{2} + \sum_{k=1}^{\infty} A_k \cos\left(\frac{2\pi kt_a}{T}\right) + B_k \sin\left(\frac{2\pi kt_a}{T}\right) \text{ όπου} \tag{1.17}$$

$$x(t_a^+) \equiv \lim_{t \to t_a^+} x(t) \text{ και } x(t_a^-) \equiv \lim_{t \to t_a^-} x(t) \tag{1.18}$$

Κάνοντας χρήση της εξίσωσης Euler

$$e^{ia} = \cos(a) + i \sin(a) \tag{1.19}$$

η εξίωση (1.15) μπορεί να γραφεί στην ισοδύναμη και πιο συμπαγή μορφή

$$x(t) = \sum_{k=-\infty}^{\infty} C_k \exp\left\{ i \frac{2\pi kt}{T} \right\} \tag{1.20}$$

όπου

$$C_0 = \frac{A_0}{2}, \quad C_{\pm k} = \frac{A_k \mp iB_k}{2} \text{ για } k \geq 1 \tag{1.21}$$

Κάνοντας χρήση των εξισώσεων (1.16α') και (1.16β') παίρνουμε

$$\begin{aligned} C_{\pm k} &= \frac{A_k \mp iB_k}{2} \\ &= \frac{1}{2} \left[\frac{2}{T} \int_{-T/2}^{T/2} x(t) \left(\cos\left(\frac{2\pi kt}{T}\right) + i \sin\left(\frac{2\pi kt}{T}\right) \right) dt \right] \\ &= \frac{1}{T} \int_{-T/2}^{T/2} x(t) \exp\left\{ \mp i \frac{2\pi kt}{T} \right\} dt \quad (\because e^{\mp ia} = \cos a \mp i \sin a) \end{aligned} \tag{1.22}$$

Τέλος, επειδή

$$C_0 = \frac{A_0}{2} = \frac{1}{2}\left[\frac{2}{T}\int_{-T/2}^{T/2} x(t)\cos 0\ dt\right] = \frac{1}{T}\int_{-T/2}^{T/2} x(t)\,dt \tag{1.23}$$

μπορούμε να χρησιμοποιήσουμε τη σχέση

$$C_k = \frac{1}{T}\int_{-T/2}^{T/2} x(t)\exp\left\{-i\frac{2\pi kt}{T}\right\}\ dt \tag{1.24}$$

για κάθε $k \in \mathbb{Z}$.

Οι συνθήκες Dirichlet δεν είναι πολύ αυστηρές όσον αφορά τη μορφή της συνάρτησης $x(t)$ με την εξαίρεση φυσικά ότι προϋποθέτει να είναι περιοδική. Από τη στιγμή όμως που επιτρέπονται ασυνέχειες στη συνάρτηση, μπορούμε να χρησιμοποιήσουμε τους ίδιους τύπους για συναρτήσεις που είναι ορισμένες μόνο σε κάποιο διάστημα του $\mathbb{R}$, υποθέτοντας ότι είναι περιοδικές μέσω μιας *περιοδικής επέκτασής* τους.

1.4.2 Μετασχηματισμός Fourier

Οι συναρτήσεις που δεν καλύπτονται από τις σειρές Fourier είναι οι απεριοδικές συναρτήσεις. Εύκολα όμως, μπορεί να γίνει μια γενίκευση θεωρώντας ότι μια απεριοδική συνάρτηση *είναι* περιοδική, με περίοδο $T \to \infty$.

Για ευκολία στο συμβολισμό, θέτουμε

$$\omega_k \equiv \frac{2\pi k}{T} \tag{1.25}$$

Η διαφορά μεταξύ δύο διαδοχικών τιμών ω_k είναι

$$d\omega_k = \frac{2\pi(dk)}{T} = \frac{2\pi}{T} \tag{1.26}$$

και από τη στιγμή που μας ενδιαφέρει το όριο $\to \infty$, έχουμε $d\omega_k \to 0$, συνεπώς η ποσότητα ω_k γίνεται συνεχής μεταβλητή. Αντικαθιστώντας στην εξίσωση (1.20) παίρνουμε

$$\begin{aligned} x(t) &= \sum_{n=-\infty}^{\infty} C_k e^{i\omega_k t}(dk) \\ &= \sum_{k=-\infty}^{\infty} C_k e^{i\omega_k t}\left(\frac{T}{2\pi}d\omega_k\right) \end{aligned} \tag{1.27}$$

και επειδή $d\omega_k \to 0$, το άθροισμα γίνεται ολοκλήρωμα:

$$\begin{aligned} x(t) &= \int_{-\infty}^{\infty} \left(C_n \frac{T}{2\pi} \right) e^{i\omega_k t}\, d\omega_k \\ &= \int_{-\infty}^{\infty} C(k_n) e^{i\omega_j t}\, d\omega_k \end{aligned} \tag{1.28}$$

όπου $C(\omega_k) \equiv (T/2\pi)C_k$. Πλέον, μπορούμε να διώξουμε το δείκτη k (είναι περιττός αφού $\omega \equiv 2\pi k/T$), ώστε να καταλήξουμε στο ζεύγος εξισώσεων του *Μετασχηματισμού Fourier* της συνάρτησης $x(t)$

$$X(e^{i\omega}) = \mathcal{F}\left\{x(t)\right\} = \frac{1}{2\pi} \int_{-\infty}^{\infty} x(t) e^{-i\omega t}\, dt \tag{1.29α'}$$

$$x(t) = \mathcal{F}^{-1}\left\{X(\omega)\right\} = \int_{-\infty}^{\infty} X(e^{i\omega}) e^{i\omega t}\, d\omega \tag{1.29β'}$$

1.4.3 Διακριτός Μετασχηματισμός Fourier

Ο Μετασχηματισμός Fourier είναι ένα σπουδαίο μαθηματικό εργαλείο και χρησιμεύει στην επίλυση πολλών προβλημάτων, αλλά το γεγονός ότι οι μεταβλητές χρόνου και συχνότητας είναι συνεχείς, τον καθιστά ακατάλληλο για χρήση στην ψηφιακή επεξεργασία σήματος. Η λύση σε αυτό το πρόβλημα δίνεται από το *Διακριτό Μετασχηματισμό Fourier (**D**iscrete **F**ourier **T**ransform - DFT)* που αποτελεί μία προσέγγιση του Μετασχηματισμού Fourier, μέσω διακριτοποίησης των χώρων χρόνου και συχνότητας.

Η πρώτη προσέγγιση γίνεται υποθέτοντας διακριτό χρόνο μέσω του *Μετασχηματισμού Fourier Διακριτού Χρόνου (**D**iscrete **T**ime **F**ourier **T**ransform - DTFT)* που δίνεται από το ζεύγος εξισώσεων

$$X(e^{i\omega}) = DTFT\{x[n]\} = \sum_{n=-\infty}^{\infty} x[n] e^{-i\omega n} \tag{1.30α'}$$

$$x[n] = DTFT^{-1}\left\{X(e^{i\omega})\right\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{i\omega}) e^{i\omega n}\, d\omega \tag{1.30β'}$$

Έχουμε ήδη αναφέρει ότι ο διακριτός χρόνος συνεπάγεται περιορισμό του εύρους συχνοτήτων, αλλά αυτό μπορούμε να το συμπεράνουμε και από την περιοδικότητα του Μετασχηματισμού Fourier Διακριτού Χρόνου αφού

$$\begin{aligned} X(\omega+2\pi) &= \sum_{n=-\infty}^{\infty} x[n]e^{-in(\omega+2\pi)} \\ &= \sum_{n=-\infty}^{\infty} x[n]e^{-in\omega}\underbrace{e^{-i2\pi n}}_{=1} \\ &= \sum_{n=-\infty}^{\infty} x[n]e^{-in\omega} \\ &= X(\omega) \end{aligned} \tag{1.31}$$

Με αυτόν τον τρόπο δικαιολογούνται και τα όρια ολοκλήρωσης της εξίσωσης (1.30β′). Δεν πρέπει όμως να ξεχάσουμε ότι η συχνότητα ω παραμένει συνεχής μεταβλητή, οπότε για να πάρουμε τις εξισώσεις του Διακριτού Μετασχηματισμού Fourier μένει η δειγματοληψία της συχνότητας. Από τη συμμετρία των μετασχηματισμών είναι εύκολο να δειχθεί ότι η διακριτοποίηση του χώρου συχνότητας οδηγεί σε περιορισμό του χρόνου. Έτσι έχει νόημα να μιλάμε για σήματα πεπερασμένης διάρκειας. Ας υποθέσουμε λοιπόν ότι η $x[n]$ είναι ορισμένη στο διάστημα $[0, N-1]$, έχουμε δηλαδή N δείγματά της. Η δειγματοληψία της συχνότητας πρέπει να γίνει με τέτοιο τρόπο ώστε να μην χάνεται πληροφορία μέσω του μετασχηματισμού. Χωρίς να δώσουμε τις σχετικές αποδείξεις, προχωράμε κατευθείαν στο αποτέλεσμα που δίνεται χωρίζοντας το χώρο συχνότητας σε N ίσα τμήματα ώστε να καταλήξουμε στο ζεύγος εξισώσεων του Διακριτού Μετασχηματισμού Fourier

$$x[n] = \frac{1}{N}\sum_{k=0}^{N-1} X[k]\exp\left\{i\frac{2\pi kn}{N}\right\}, \quad 0 \le n \le N-1 \tag{1.32α′}$$

$$X[k] = \sum_{n=0}^{N-1} x[n]\exp\left\{-i\frac{2\pi kn}{N}\right\}, \quad 0 \le k \le N-1 \tag{1.32β′}$$

Τέλος, ο *Μετασχηματισμός Fourier Βραχέος Χρόνου - **S**hort **T**ime **F**ourier **T**ransform - STFT* στη διακριτή περίπτωση, δεν είναι παρά ο Διακριτός Μετασχηματισμός Fourier ενός μικρού μέρους του σήματος που προσδιορίζεται από μια παραθυρική συνάρτηση

(δηλ. ενός πλαισίου) όπως είναι αναμενόμενο στα πλαίσια της Βραχύχρονης Ανάλυσης, δηλαδή

$$STFT\{x[n]\}(m,\omega) = X(m,\omega) = \sum_{n=\infty}^{\infty} x[n]w[n-m]e^{-i\omega n} \tag{1.33}$$

Σημείωση 1. *Όλες οι μορφές του Μετασχηματισμού Fourier υποθέτουν μιγαδικά σήματα. Στην πράξη βέβαια, στην επεξεργασία ομιλίας (αλλά και στους περισσότερους τομείς της Ανάλυσης Σήματος) χρησιμοποιούμε πραγματικά σήματα. Παρατηρούμε τώρα ότι στην εξίσωση (1.32α') οι τιμές $X[0]$ και $X[N/2]$[2] είναι πραγματικές. Επίσης, για πραγματικά σήματα $x[n]$ έχουμε*

$$\begin{aligned}
X[N-k] &= \sum_{n=0}^{N-1} x[n]\exp\left\{-i\,\frac{2\pi(N-k)n}{N}\right\} \\
&= \sum_{n=0}^{N-1} x[n]\left(\cos\left(\frac{2\pi(N-k)n}{N}\right) - i\,\sin\left(\frac{2\pi(N-k)n}{N}\right)\right) \\
&= \sum_{n=0}^{N-1} x[n]\left(\cos\left(-\frac{2\pi kn}{N}\right) - i\,\sin\left(-\frac{2\pi kn}{N}\right)\right) \\
&= \sum_{n=0}^{N-1} x[n]\left(\cos\left(\frac{2\pi kn}{N}\right) + i\,\sin\left(\frac{2\pi kn}{N}\right)\right) \\
&= X^*[k]
\end{aligned} \tag{1.34}$$

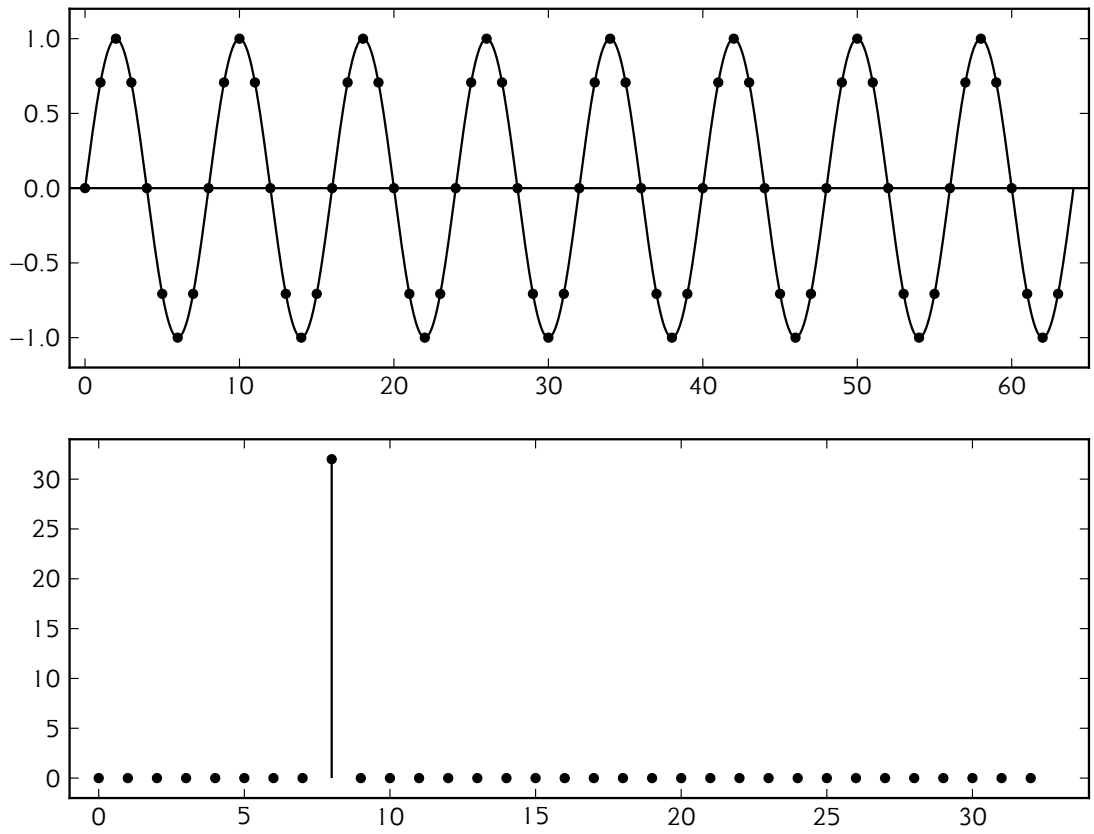
όπου με $X^*[k]$ συμβολίζουμε το μιγαδικό συζυγή του $X[k]$. Συνεπώς, εκτός από τις δύο πραγματικές τιμές χρειαζόμαστε μόνο $N/2-1$ μιγαδικές τιμές για να προσδιορίσουμε όλο το φάσμα. Επίσης, θεωρείται ότι το πλάτος του μετασχηματισμού περιέχει το μεγαλύτερο μέρος της πληροφορίας για τα σήματα ομιλίας [Anderson, 1984], οπότε παραδοσιακά η φάση τις περισσότερες φορές αγνοείται. Αυτό σημαίνει ότι για ένα μετασχηματισμό μήκους N χρησιμοποιούνται $N/2+1$ στοιχεία και αυτό θα υιοθετήσουμε από εδώ και στο εξής. Σημειώνουμε πάντως, ότι αγνοώντας τη φάση δεν μπορούμε να επιστρέψουμε στην αρχική μορφή και έτσι, σε εφαρμογές όπου απαιτείται η ανακατασκευή του σήματος (π.χ. ενίσχυση σήματος ομιλίας, σύνθεση ομιλίας κτλ.) αποθηκεύονται και οι N τιμές.

[2]Προφανώς, αυτό ισχύει στην περίπτωση όπου το μήκος N είναι άρτιος αριθμός. Στην περίπτωση περιττού μήκους η μόνη πραγματική τιμή του μετασχηματισμού είναι η $X[0]$.

Ο Μετασχηματισμός Fourier Βραχέος Χρόνου έχει αναδειχθεί σε ένα από τα σημαντικότερα εργαλεία επεξεργασίας σήματος ομιλίας. Παρ' όλα αυτά, η διακριτοποίηση έχει κάποιες συνέπειες που πρέπει να ληφθούν υπόψη κατά τη χρήση του. Ας θυμηθούμε, ότι για ένα πλαίσιο μήκους N και με ρυθμό δειγματοληψίας f_s, ο Διακριτός Μετασχηματισμός Fourier υπολογίζεται στις συχνότητες

$$f_k = \frac{k f_s}{N}, \quad k = 0, 1, 2, \ldots, N-1 \tag{1.35}$$

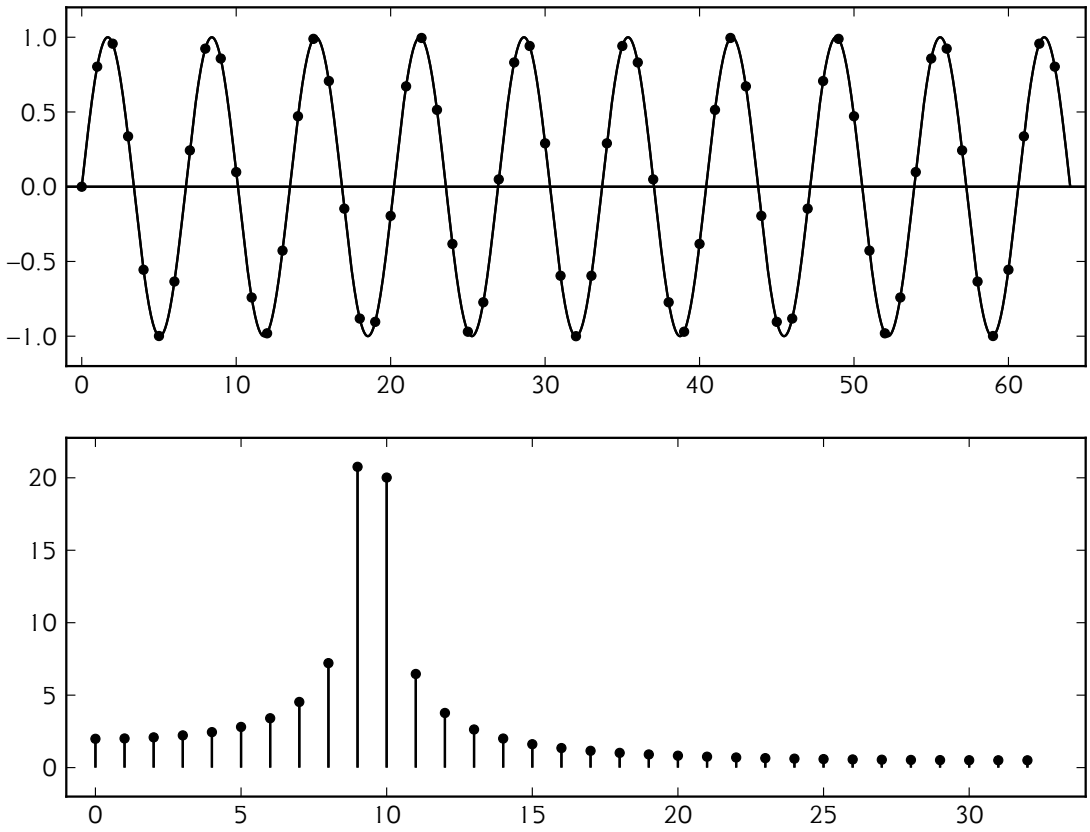
Στο Σχήμα 1.16 έχουμε 8 περιόδους ενός ημιτόνου και εκτελούμε Διακριτό Μετασχηματισμό Fourier μήκους $N = 64$. Το αποτέλεσμα είναι το αναμενόμενο, αφού στο φάσμα παρατηρείται μόνο μία κορυφή στη συχνότητα $k = 8$ και οι υπόλοιπες τιμές είναι μηδενικές.



Σχήμα 1.16

Αντίθετα, στο Σχήμα 1.17 έχουμε 9.5 περιόδους του ίδιου ημιτόνου. Η τιμή

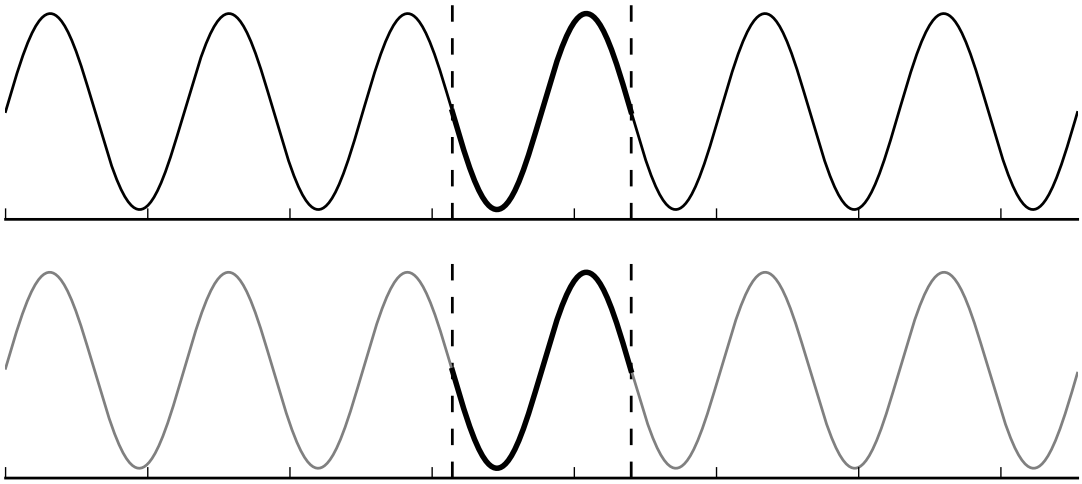
9.5 δεν μπορεί να αναπαρασταθεί στο διακριτό χώρο συχνότητας και η μορφή του μετασχηματισμού είναι πολύ διαφορετική, αφού έχουμε μη-μηδενικές τιμές σε όλο το μήκος του φάσματος. Αυτό το φαινόμενο ονομάζεται *φασματική διαρροή*, περιγράφοντας μια εικόνα όπου η ενέργεια που περιέχεται σε μια συχνότητα "διαρρέει" στις γύρω της.



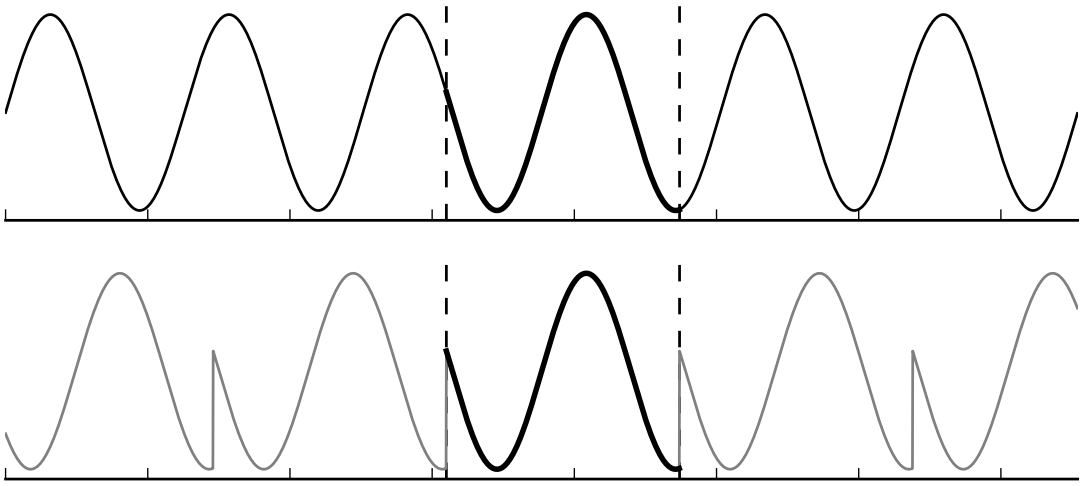
Σχήμα 1.17

Ο παράγοντας που προκαλεί τη φασματική διαρροή, δεν είναι η δειγματοληψία, αλλά το χρονικά περιορισμένο σήμα που επιλέγεται από την παραθυρική συνάρτηση $w[n]$. Πρέπει να θυμηθούμε ότι σε κάθε περίπτωση, ο μετασχηματισμός "βλέπει" μια περιοδική επέκταση του επιλεγμένου τμήματος. Έτσι, εάν το παράθυρο είναι τοποθετημένο ώστε να περιέχει *ακριβώς* ακέραιο αριθμό περιόδων, η περιοδική επέκταση δεν έχει διαφορά από το αρχικό σήμα (Σχήμα 1.18) και το αποτέλεσμα είναι ακριβές. Από την άλλη, αν δεν υπάρχει ακέραιος αριθμός περιόδων στο επιλεγμένο τμήμα, δη-

μιουργούνται ασυνέχειες στην περιοδική επέκταση (Σχήμα 1.19) και έχουμε διαρροή.



Σχήμα 1.18: Επιλογή και περιοδική επέκταση τμήματος με ακέραιο αριθμό περιόδων



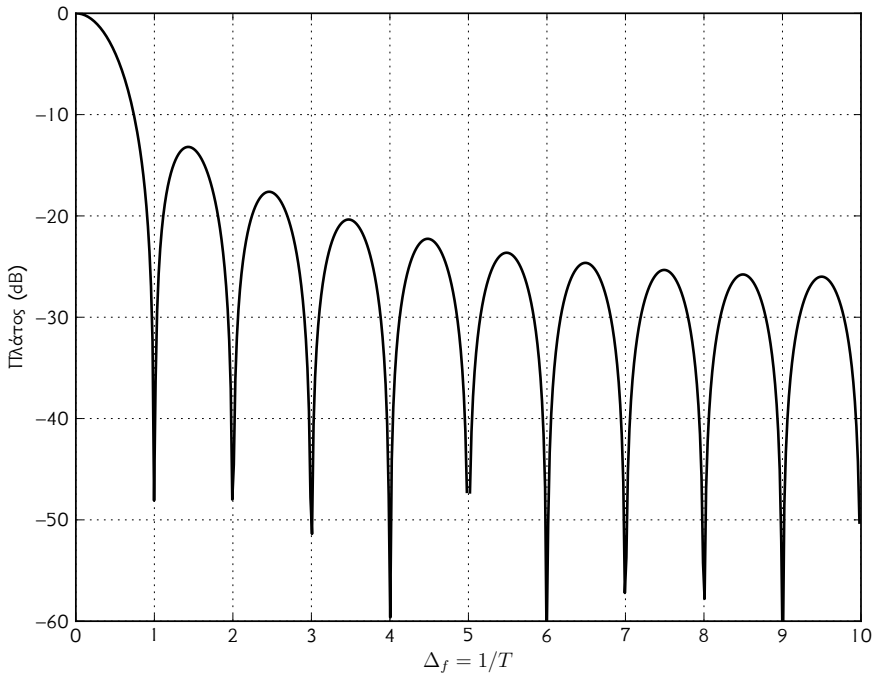
Σχήμα 1.19: Επιλογή και περιοδική επέκταση τμήματος με μη ακέραιο αριθμό περιόδων

Με άλλα λόγια, η αιτία της φασματικής διαρροής εντοπίζεται *στο παράθυρο* $w[n]$ που δημιουργεί μεγάλες ασυνέχειες στα άκρα του τμήματος όπου τοποθετείται. Το πρόβλημα αυτό περιορίζεται από τη χρήση παραθύρων που δεν είναι σταθερά όπως

στην εξίσωση (1.7) αλλά φθίνουν ομαλά προς το 0 και από τις δύο πλευρές, προκαλώντας έτσι μικρότερες ασυνέχειες. Για να δούμε πως επηρρεάζεται το φάσμα από το σχήμα του παραθύρου σημειώνουμε ότι όπως και σε άλλους ολοκληρωτικούς μετασχηματισμούς, έτσι και στο Μετασχηματισμό Fourier ισχύει το Θεώρημα Συνέλιξης, δηλαδή

$$\mathcal{F}\{x(t)w(t)\} = \mathcal{F}\{x(t)\} * \mathcal{F}\{w(t)\} = X(\omega) * W(\omega) \tag{1.36}$$

Στο Σχήμα 1.20 βλέπουμε το πλάτος του Διακριτού Μετασχηματισμού Fourier του παραθύρου της εξίσωσης (1.7), γνωστό και ως *ορθογώνιο παράθυρο* Οι λοβοί που



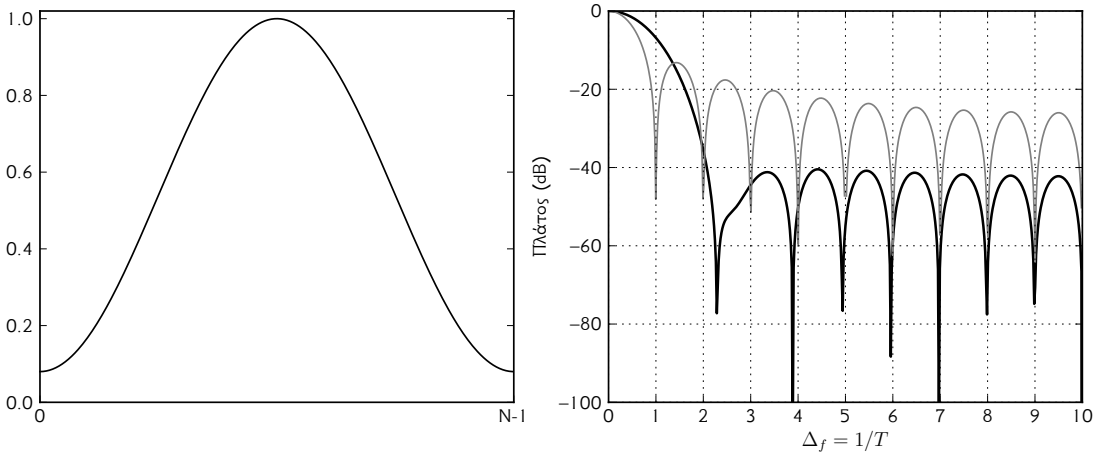
Σχήμα 1.20

δημιουργούνται εξηγούν το φαινόμενο της φασματικής διαρροής κοιτώντας το από τη συχνοτική διάσταση. Γενικά είναι επιθυμητό να έχουμε ένα στενό κύριο λοβό (το λοβό γύρω από το 0) ώστε να υπάρχει ικανότητα διάκρισης μεταξύ κοντινών συχνοτήτων, αλλά και μικρό πλάτος των δευτερευόντων λοβών. Αν και υπάρχουν τρόποι να μετρηθούν οι ιδιότητες ενός παραθύρου, θα αποφύγουμε την παράθεση των ποσοτικών χαρακτηριστικών και θα τις αναπαραστήσουμε σχηματικά, χρησιμοποιώντας το

ορθογώνιο παράθυρο σαν μέτρο σύγκρισης. Μερικά από τα πιο συνήθη παράθυρα είναι τα εξής:

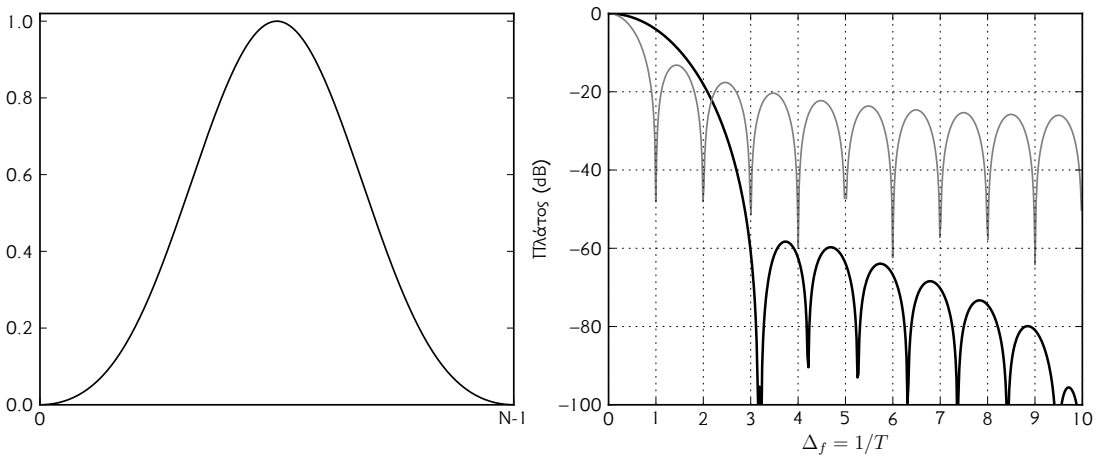
Παράθυρο Hamming

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1 \tag{1.37}$$



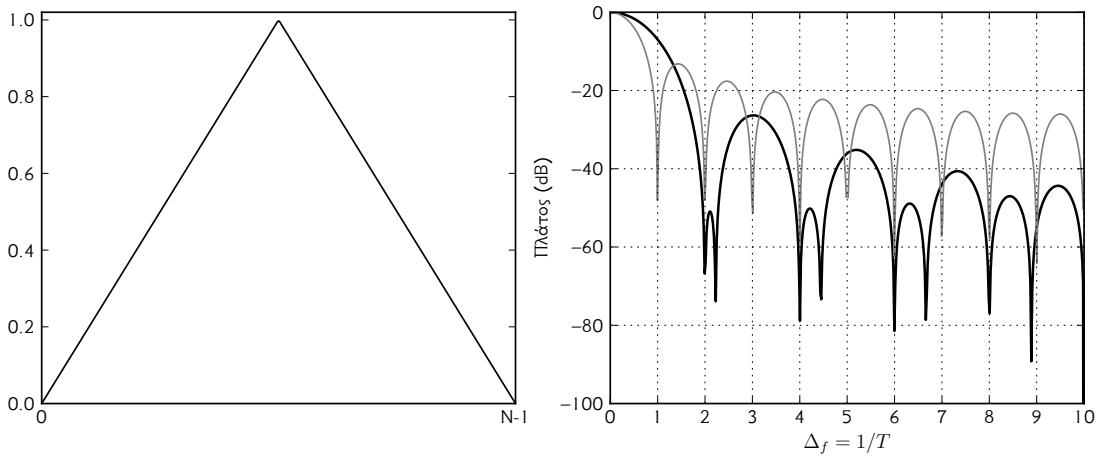
Παράθυρο Blackman

$$w[n] = 0.42 - 0.5\cos\left(\frac{2\pi n}{N}\right) + 0.08\cos\left(\frac{4\pi n}{N}\right), \quad 0 \le n \le N \tag{1.38}$$



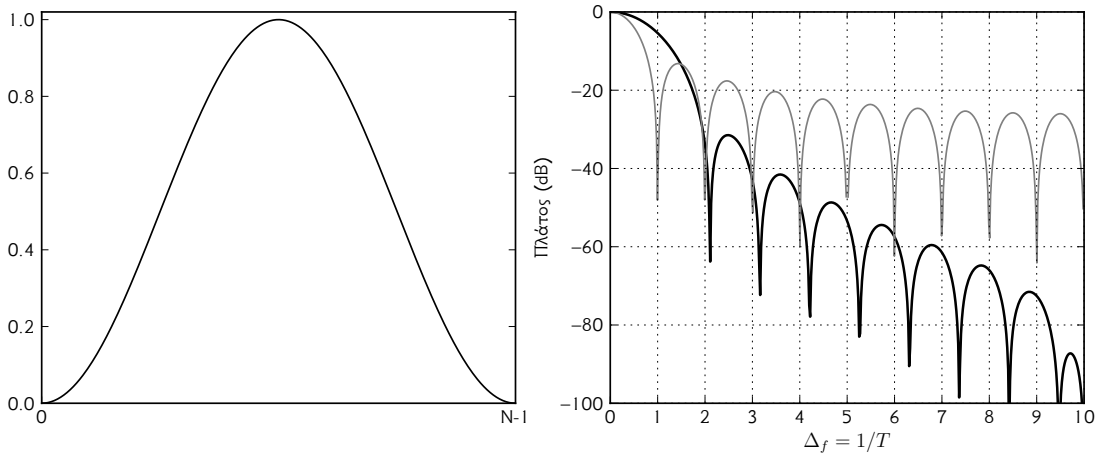
Παράθυρο Bartlett

$$w[n] = 1 - \left|\frac{2n}{N-1} - 1\right|, \quad 0 \le n \le N-1 \tag{1.39}$$



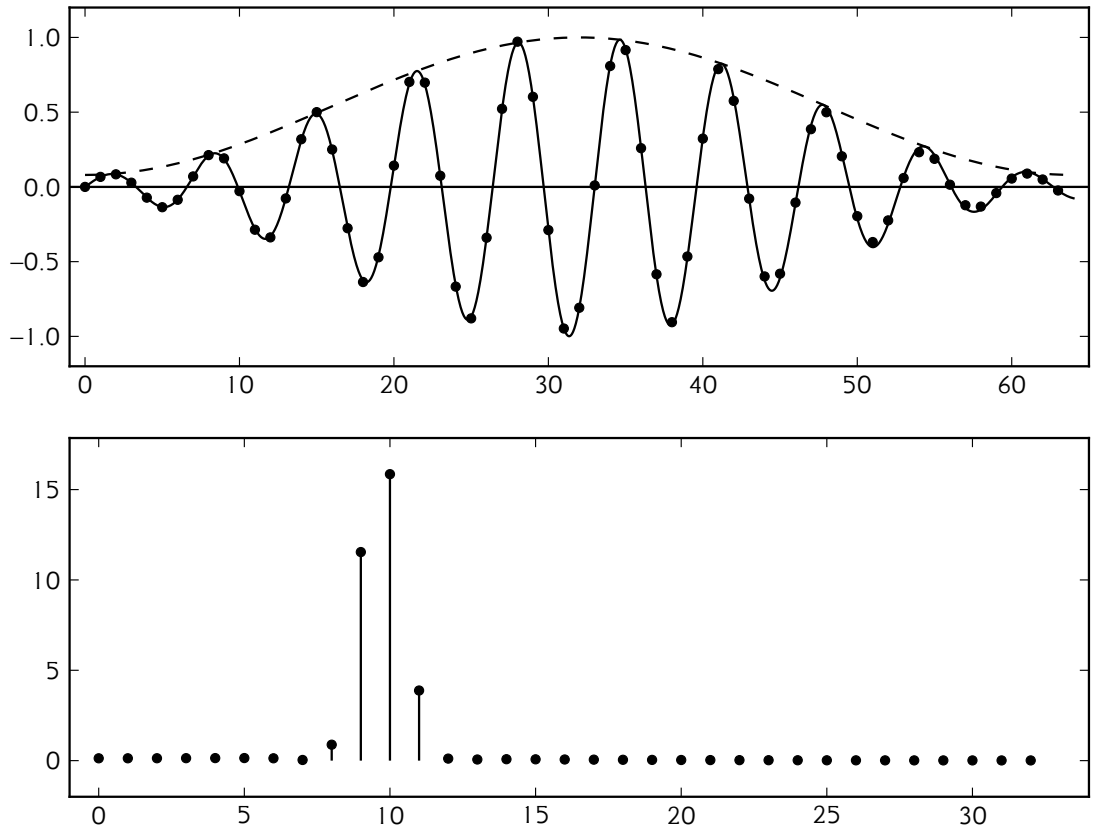
Παράθυρο Hann (ή Hanning)

$$0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1 \tag{1.40}$$



Όπως μπορούμε να δούμε, η επιλογή ενός παραθύρου είναι ένα είδος "συμβιβασμού" ανάμεσα στο εύρος του κύριου λοβού και το πλάτος των δευτερευόντων και στην

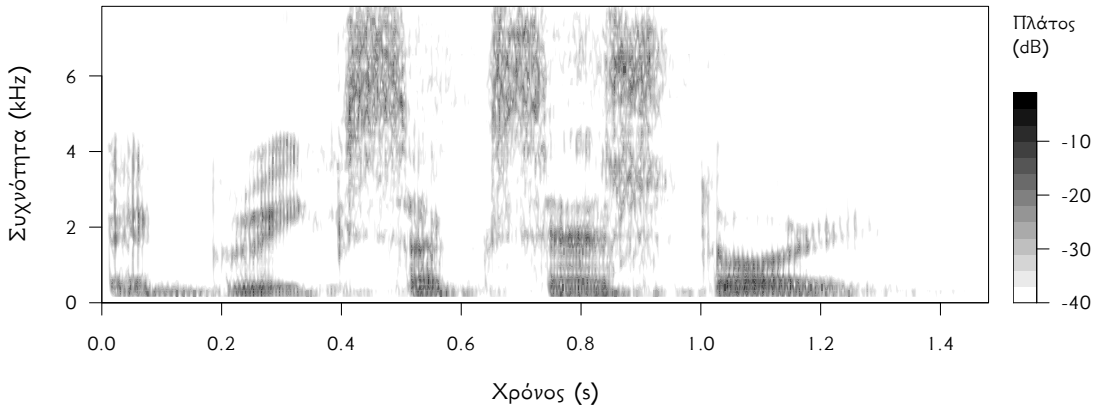
πράξη γίνεται σύμφωνα με την εφαρμογή. Όσον αφορά την επεξεργασία ομιλίας, οι πιο δημοφιλείς επιλογές είναι τα παράθυρα Hamming και Hann. Στο Σχήμα 1.21 χρησιμοποιούμε το σήμα του Σχήματος 1.17 και το παράθυρο Hamming. Η μείωση της φασματικής διαρροής είναι εμφανής.



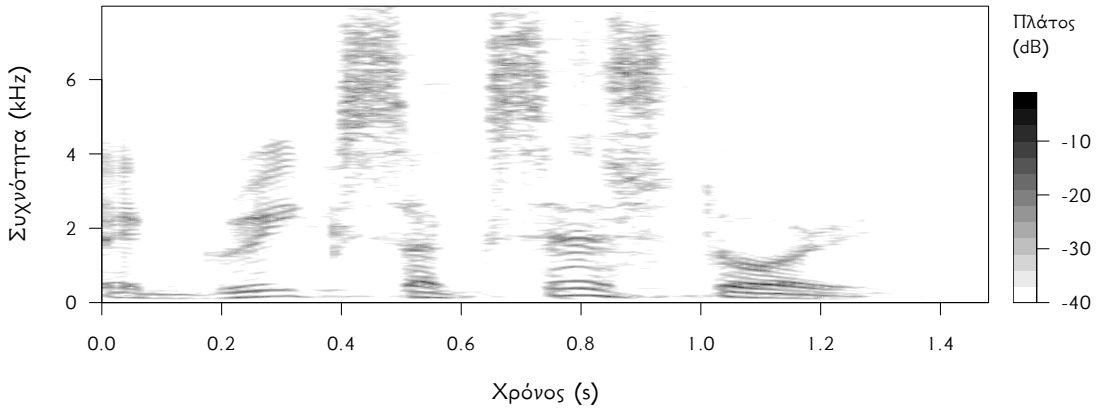
Σχήμα 1.21

Αυτός δεν είναι ο μόνος συμβιβασμός που γίνεται στα πλαίσια της βραχύχρονης επεξεργασίας. Κάτι άλλο που θα πρέπει επίσης να προσέξουμε είναι *το μήκος* του παραθύρου. Από την εξίσωση (1.33) βλέπουμε ότι ο Μετασχηματισμός Fourier ενός σήματος είναι συνάρτηση της συχνότητας αλλά *και* του χρόνου, υπό την έννοια ότι επιλέγουμε τη χρονική τοποθέτηση του παραθύρου. Το μήκος του N όμως είναι και το μήκος του μετασχηματισμού. Έτσι, εάν επιθυμούμε καλή *χρονική ανάλυση* και χρησιμοποιήσουμε ένα μικρό παράθυρο (εδώ ισχύουν οι ίδιες ιδέες που αναφέραμε στην προηγούμενη ενότητα, βλ. Σχήμα 1.15) καταλήγουμε με χαμηλή *συχνοτική ανάλυση* και

το αντίστροφο. Αυτό μπορεί να γίνει πιο κατανοητό μέσω των φασματογραμμάτων των Σχημάτων 1.22 και 1.23.



Σχήμα 1.22: Φασματόγραμμα σήματος ομιλίας με μικρό παράθυρο ($f_s = 16$ kHz, $N = 100$



Σχήμα 1.23: Φασματόγραμμα σήματος ομιλίας με μεγάλο παράθυρο ($f_s = 16$ kHz, $N = 640$

Το φασματόγραμμα είναι μία ψεύδο-τρισδιάστατη απεικόνηση του πλάτους του Μετασχηματισμού Fourier Βραχέος Χρόνου, όπου ο ένας άξονας αντιστοιχεί στο χρόνο, ο άλλος στη συχνότητα και το τετράγωνο του πλάτους (φάσμα ισχύος) περιγράφεται με χρώματα ή επίπεδα του γκρι. Παρατηρούμε ότι για μικρό N σχηματίζονται κάθετες γραμμές που υποδηλώνουν τη χαμηλή διακριτικότητα ως προς τη συχνότητα, ενώ για

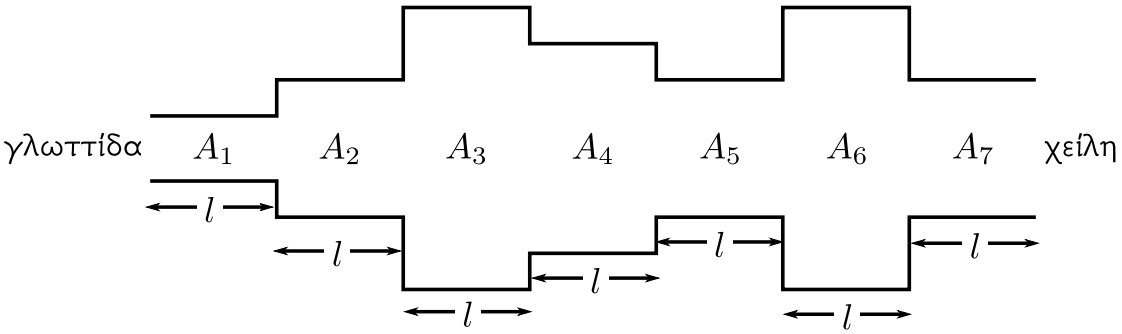
μεγάλο N οριζόντιες που αντιστοιχούν σε χαμηλή διακριτικότητα ως προς το χρόνο. Και πάλι όμως δεν υπάρχει "χρυσός κανόνας" για την επιλογή μήκους παραθύρου, αλλά υπολογίζεται σύμφωνα με τις ανάγκες της εφαρμογής.

1.5 Γραμμική Πρόβλεψη

Η μέθοδος της ανάλυσης γραμμικής πρόβλεψης αποτελεί μια από τις κύριες τεχνικές για την εκτίμηση βασικών παραμέτρων ομιλίας, όπως τονικό ύψος, φωνοσυντονισμοί, συναρτήσεις φωνητικής οδού, αλλά και για την αναπαράσταση του ήχου για μετάδοση ή αποθήκευση με χαμηλό κόστος. Η αξία της μεθόδου αυτής βρίσκεται τόσο στην ικανότητά της να παρέχει πολύ ακριβείς προσεγγίσεις των διάφορων παραμέτρων, όσο και στη σχετικά μικρή πολυπλοκότητά της.

1.5.1 Μοντέλο ιδανικών σωλήνων

Η φιλοσοφία της γραμμικής πρόβλεψης συνδέεται στενά με το μοντέλο παραγωγής ομιλίας. Συγκεκριμένα, η φωνητική οδός μοντελοποιείται ως μία σειρά κυλίνδρων ίσου μήκους με ιδανικά τοιχώματα, με την έννοια ότι οι ανακλάσεις σε αυτά δεν οδηγούν σε απώλεια ενέργειας (Σχήμα 1.24). Εξαιτίας της διαφοράς των εμβαδών διατομής των



Σχήμα 1.24: Μοντελοποίηση της φωνητικής οδού ως μία σειρά ιδανικών σωλήνων

σωλήνων, όταν ένα ηχητικό κύμα οδεύει από τη γλωττίδα προς τα χείλη, σε κάθε ένωση σωλήνων ένα τμήμα του ανακλάται προς τα πίσω και το υπόλοιπο συνεχίζει. Η k-οστή ένωση του μοντέλου λοιπόν, χαρακτηρίζεται από ένα *συντελεστή ανάκλασης*

(εξίσωση 1.41α′) ή ισοδύναμα από ένα *συντελεστή διάδοσης* (εξίσωση 1.41β′), ή ακόμα από το λόγο των δύο σωλήνων εκατέρωθέν της A_k/A_{k+1}

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \tag{1.41α′}$$

$$t_k = 1 - r_k = \frac{2A_k}{A_{k+1} + A_k} \tag{1.41β′}$$

Κάνοντας μία σχετικά απλή ανάλυση της ροής του αέρα μέσα σε ένα μοντέλο ιδανικών σωλήνων [Rabiner and Schafer, 1978], φτάνουμε στο συμπέρασμα ότι η συνάρτηση μεταφοράς του μπορεί πάντα να εκφραστεί ως

$$V(z) = \frac{0.5(1 + r_G)\prod_{k=1}^{p}(1 + r_k)z^{-p/2}}{A(z)} \tag{1.42}$$

όπου r_G είναι ο συντελεστής ανάκλασης στη γλωττίδα και $A(z)$ ένα πολυώνυμο του z^{-1} της μορφής

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} \tag{1.43}$$

Με άλλα λόγια, η συνάρτηση μεταφοράς του μοντέλου ιδανικών σωλήνων περιέχει έναν παράγοντα χρονικής υστέρησης που αντιστοιχεί στον αριθμό των σωλήνων και έχει *μόνο πόλους* οι οποίοι ορίζουν τις συχνότητες συντονισμού του μοντέλου. Αυτή η παρατήρηση είναι η βάση της Ανάλυσης Γραμμικής Πρόβλεψης: Θεωρούμε ότι έχουμε ένα σύστημα με συνάρτηση μεταφοράς

$$H(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{1.44}$$

Τότε η ακολουθία εξόδου $x[n]$ συνδέεται με τη διέγερση $u[n]$ από τη σχέση

$$x[n] = \sum_{k=1}^{p} x[n-k] + Gu[n] \tag{1.45}$$

Η σύνδεσή του με το μοντέλο "πηγής-φίλτρου" του Σχήματος 1.10 είναι προφανής.

1.5.2 Εκτίμηση Παραμέτρων

Το όνομα "γραμμική πρόβλεψη" δικαιολογείται από τη σχέση

$$\tilde{x}[n] = \sum_{k=1}^{p} a_k x[n-k] \tag{1.46}$$

όπου ένα δείγμα εκτιμάται ως γραμμικός συνδυασμός προηγούμενων δειγμάτων. Μια ισοδύναμη διατύπωση είναι ότι εφαρμόζουμε ένα *αυτοπαλινδρομικό μοντέλο* [Brockwell and Davis, 1986] στο υπό εξέταση πλαίσιο ομιλίας. Η εκτίμηση των συντελεστών a_k γίνεται μέσω της ελαχιστοποίησης του *σφάλματος πρόβλεψης βραχέος χρόνου*

$$\begin{aligned} E &= \sum_{n=0}^{N+p-1} e^{[}n] \\ &= \sum_{n=0}^{N+p-1} (x[n] - \tilde{x}[n]) \\ &= \sum_{n=0}^{N+p-1} \left(x[n] - \sum_{k=1}^{p} a_k x[n-k] \right)^2 \end{aligned} \tag{1.47}$$

Τα όρια άθροισης στην ουσία δηλώνουν τη χρήση ενός βραχύχρονου παραθύρου ανάλυσης. Με αυτό τον τρόπο, έχουμε $x[n] \equiv 0$ εκτός του διαστήματος $[0, N-1]$ και συνεπώς μηδενικό σφάλμα πρόβλεψης εκτός του διαστήματος $[0, N+p-1]$. Θέτοντας τώρα $\partial E/\partial a_i = 0$ για $i = 1, 2, \ldots, p$, παίρνουμε

$$\sum_{n=0}^{N+p-1} x[n-i]x[n] = \sum_{k=1}^{p} \hat{a}_k \sum_{n=0}^{N+p-1} x[n-i]x[n-k], \quad 1 \le i \le p \tag{1.48}$$

Κάνοντας χρήση της δειγματικής συνάρτησης αυτοσυσχέτισης

$$R[k] = \sum_{m=0}^{N-1-k} x[m]x[m+k] \tag{1.49}$$

μπορούμε να γράψουμε τις εξισώσεις 1.48 ως

$$\sum_{k=1}^{p} \hat{a}_k R[i-k] = R[i], \quad 1 \le i \le p \tag{1.50}$$

ή και σε μορφή πινάκων

$$\begin{bmatrix} R[0] & R[1] & R[2] & \cdots & R[p-1] \\ R[1] & R[0] & R[1] & \cdots & R[p-2] \\ R[2] & R[1] & R[0] & \cdots & R[p-3] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R[p-1] & R[p-2] & R[p-3] & \cdots & R[0] \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \vdots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} R[1] \\ R[2] \\ R[3] \\ \vdots \\ R[p] \end{bmatrix} \tag{1.51}$$

Το σύστημα των εξισώσεων 1.51 μπορεί να λυθεί με όλους τους συνήθεις τρόπους, αλλά εξαιτίας του συμμετρικού Toeplitz [3] πίνακα, συνήθως χρησιμοποιούνται πιο εξειδικευμένοι αλγόριθμοι όπως ο Levinson-Durbin (Αλγόριθμος 1).

Αλγόριθμος 1 Επίλυση συστήματος Toeplitz [Levinson, 1947], [Durbin, 1960]

1: Αρχικοποίηση

$$E^{(0)} = \gamma[0] \tag{1.52}$$

2: Επανάληψη για $i = 1, \ldots, p$

$$k_i = \frac{1}{E^{(i-1)}} \left(\gamma[i] - \sum_{j=1}^{i-1} a_j^{(i-1)} \gamma[i-j] \right) \tag{1.53α'}$$

$$a_i^{(i)} = k_i \tag{1.53β'}$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}, \quad 1 \leq j < i \tag{1.53γ'}$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)} \tag{1.53δ'}$$

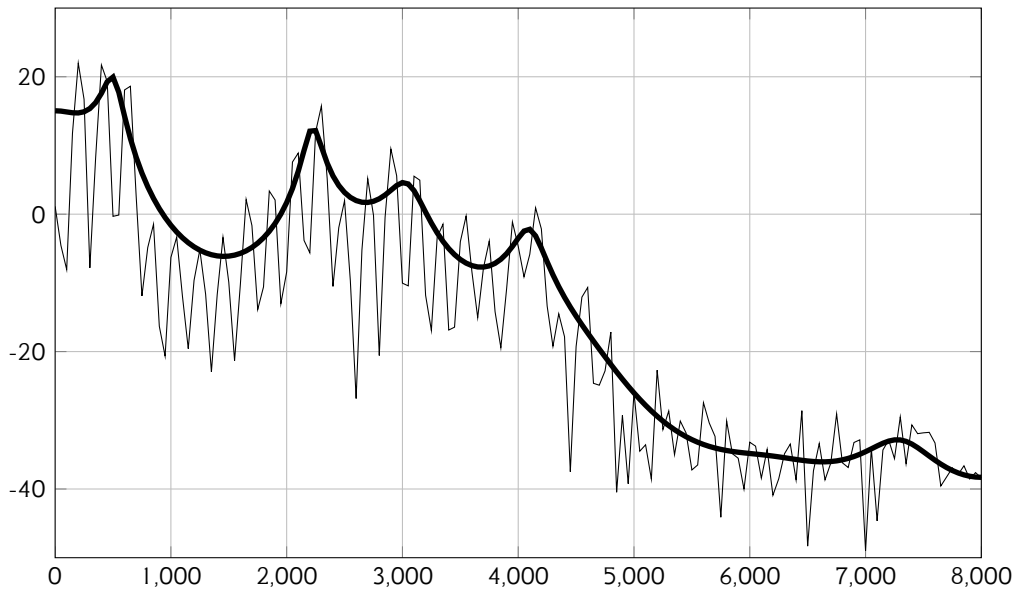
3: Τελική λύση

$$a_j = a_j^{(p)}, \quad 1 \leq j \leq p \tag{1.54}$$

Σύμφωνα με όσα έχουμε υποθέσει για τον τρόπο παραγωγής ομιλίας αλλά και το μοντέλο ιδανικών σωλήνων, το σήμα που εκτιμάται μέσω γραμμικής πρόβλεψης αντανακλά τις ιδιότητες της φωνητικής οδού. Στο επίπεδο συχνότητας, από τη στιγμή που η συνάρτηση μεταφοράς της φωνητικής οδού έχει μόνο πόλους, μέσω της εκτίμησής μας προσεγγίζουμε *τις κορυφές* του φάσματος, σχηματίζοντας έτσι το *φασματικό φάκελο* (Σχήμα 1.25).

Η έλλειψη μηδενικών στη συνάρτηση μεταφοράς έχει σαν συνέπεια την αδυναμία προσέγγισης των κοιλάδων του φάσματος. Για να λυθεί αυτό το πρόβλημα μπο-

[3] Ένας τετραγωνικός $n \times n$ πίνακας $A = \{a_{i,j}\}$ ονομάζεται Toeplitz αν τα στοιχεία κάθε διαγωνίου του είναι ίσα, ισχύει δηλαδή $a_{i,j} = a_{i+1,j+1}$ για κάθε $i, j < n$



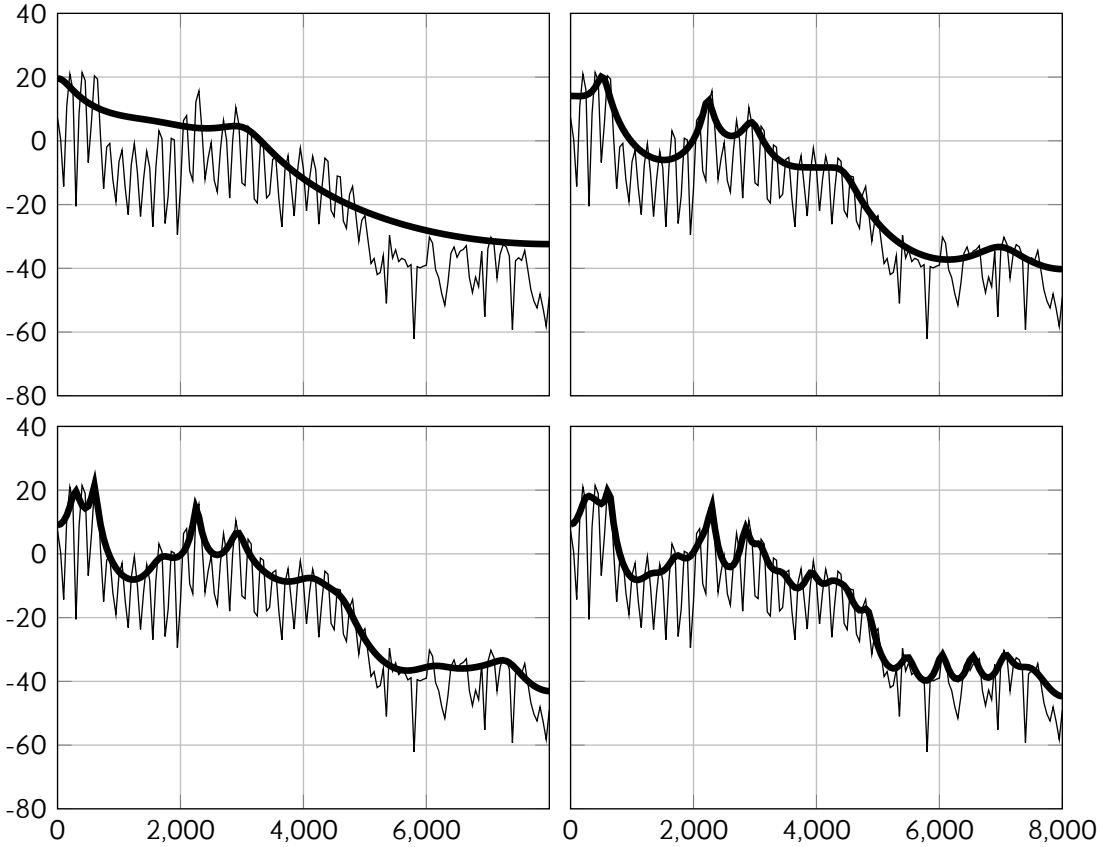
Σχήμα 1.25

ρούμε να υποθέσουμε μια εναλλακτική μοντελοποίηση της φωνητικής οδού όπου θα λαμβάνεται υπόψη και η ρινική κοιλότητα και θα οδηγούσε σε ένα μοντέλο $ARMA$ για το σήμα ομιλίας στο οποίο όμως η εκτίμηση των παραμέτρων θα ήταν πολύ πιο περίπλοκη. Τέτοια μοντέλα χρησιμοποιούνται αρκετά συχνά στη σύνθεση φωνής, αλλά στην ανάλυση όπου προτιμώνται πιο απλές προσεγγίσεις, αυτή η αδυναμία μπορεί να αντισταθμιστεί από την επιλογή μεγαλύτερης τιμής p όπως μπορούμε να δούμε στο Σχήμα 1.26

Τα μέχρι τώρα αποτελέσματα είναι από μόνα τους πολύ χρήσιμα και χρησιμοποιούνται συχνά στην αναγνώριση ομιλίας αφού οι κορυφές που αντιστοιχούν στις συχνότητες συντονισμού της φωνητικής οδού αποτελούν ένα από τα κύρια χαρακτηριστικά του πλαισίου ομιλίας. Όμως εξίσου σημαντικό είναι και το σήμα σφάλματος $e[n]$ που σχετίζεται με τη διέγερση $u[n]$ μέσω της σχέσης

$$e[n] = x[n] - \tilde{x}[n] = x[n] - \sum_{k=1}^{p} x[n-k] = Gu[n] \tag{1.55}$$

Όπως είναι αναμενόμενο, για ένα πλαίσιο έμφωνης ομιλίας θα παρουσιάζει κάποια περιοδικότητα από την οποία μπορεί να εκτιμηθεί και το τονικό ύψος της φωνής, ενώ για άφωνη ομιλία θα προσεγγίζεται από ένα μοντέλο χρονοσειράς λευκού θορύβου



Σχήμα 1.26: Φασματικός φάκελος για διάφορες τιμές p. Από αριστερά προς τα δεξιά και από πάνω προς τα κάτω, $p = 5, 16, 24, 40$

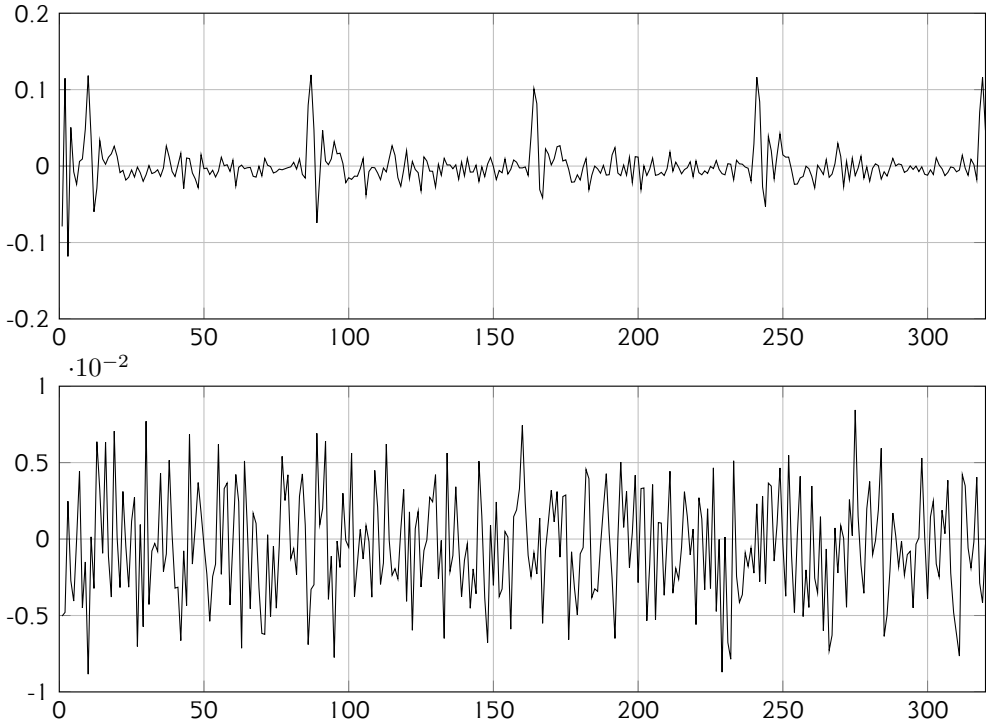
(Σχήμα 1.27).

Τέλος, μαζί με την εκτίμηση για το συντελεστή κέρδους G

$$G^2 = R[0] - \sum_{k=1}^{p} \hat{a}_k R[k] \tag{1.56}$$

που "ταιριάζει" τις ενέργειες των σημάτων, έχουμε εκτιμήσεις για *όλες* τις παραμέτρους του μοντέλου πηγής-φίλτρου.

Ουσιαστικά, μέσω της Ανάλυσης Γραμμικής Πρόβλεψης επιτυγχάνεται ο διαχωρισμός πηγής και φίλτρου και οι ιδιότητες του καθενός μπορούν να μελετηθούν απομονωμένα. Η παράμετρος που καθορίζει αυτόν τον διαχωρισμό είναι η τάξη του μοντέλου (ισοδύναμα, ο αριθμός ιδανικών σωλήνων) p, η οποία ιδανικά θα επιτρέπει στο εκτι-



Σχήμα 1.27: Σήμα σφάλματος Ανάλυσης Γραμμικής Πρόβλεψης για πλαίσιο έμφωνης (πάνω) και άφωνης (κάτω) ομιλίας

μώμενο σήμα να περιέχει ιδιότητες *μόνο* της φωνητικής οδού και στο σήμα σφάλματος ιδιότητες της πηγής. Πειραματικά έχει δειχθεί ότι ο αριθμός φωνοσυντονισμών είναι περίπου 1/kHz για τη φωνή ενός ενήλικα άνδρα και 1.3/kHz για τη φωνή μιας γυναίκας ή ενός παιδιού και από τη στιγμή που χρειάζονται 2 πόλοι ανά συχνότητα φωνοσυντονισμού, ένας εμπειρικός κανόνας για την επιλογή είναι

$$p \approxeq \frac{f_s}{1000} + 2 \tag{1.57}$$

1.5.3 Ισοδύναμες μορφές

Οι συντελεστές $\hat{a}_k$ ονομάζονται *συντελεστές πρόβλεψης (prediction coefficients)* ή απλά *συντελεστές LPC* και οι ενδιάμεσοι συντελεστές k_i στον Αλγόριθμο 1 είναι οι *συντελεστές ανάκλασης* του μοντέλου ιδανικών σωλήνων που είχαμε υποθέσει. Εναλλακτικά, χρησιμοποιώντας όρους της Ανάλυσης Χρονοσειρών, οι συντελεστές $-k_i$ είναι οι *συντελεστές*

μερικής αυτοσυσχέτισης [4] [Brockwell and Davis, 1986].

Σε κάθε περίπτωση ,τα δύο σύνολα συντελεστών είναι ισοδύναμα αφού μπορούμε να πάμε από τη μία μορφή στην άλλη μέσω της απλής αναδρομικής σχέσης

$$\begin{aligned} k_i &= \hat{a}_i^{(i)} \\ a_j^{(i-1)} &= \frac{a_j^{(i)} - a_i^{(i)} a_{i-j}^{(i)}}{1 - k_i^2}, \quad 1 \le j \le i-1 \end{aligned} \tag{1.58}$$

με το δείκτη i να παίρνει τις τιμές $p, p-1, \ldots, 1$ (με αυτή τη σειρά). Υπάρχουν πολλά μαθηματικά ισοδύναμα σύνολα συντελεστών που περιέχουν τα αποτελέσματα της ανάλυσης γραμμικής πρόβλεψης και συχνά προτιμώνται αντί των συντελεστών πρόβλεψης για διάφορες εφαρμογές (π.χ. κωδικοποίηση χαμηλού ρυθμού μετάδοσης), καθώς οι ιδιότητές τους ως προς τον κβαντισμό κτλ. είναι καλύτερες, ή αποτελούν καλύτερα χαρακτηριστικά για ορισμένες εργασίες. Κάποια παραδείγματα είναι τα εξής:

Συχνότητες LSF Οι συχνότητες LSF (**L**ine **S**pectral **F**requencies) ή ζεύγη LSP (**L**ine **S**pectral **P**airs), προτάθηκαν σαν μία πιο κατάλληλη για μετάδοση αναπαράσταση των συντελεστών πρόβλεψης στο [Itakura, 1975]. Για τον υπολογισμό τους, εκφράζουμε το πολυώνυμο γραμμικής πρόβλεψης της εξίσωσης (1.43) ως ημιάθροισμα ενός παλινδρομικού και ενός αντιπαλινδρομικού πολυωνύμου [5]

$$A(z) = 1 - \sum_{k=1}^{p} a_k z^{-k} = 0.5\,[P(z) + Q(z)] \tag{1.59}$$

όπου

$$P(z) = A(z) + z^{-p-1} A(z^{-1}) \tag{1.60α'}$$

$$P(z) = A(z) - z^{-p-1} A(z^{-1}) \tag{1.60β'}$$

Η φυσική ερμηνεία αυτών των πολυωνύμων είναι ότι το $P(z)$ αντιστοιχεί στην κατάσταση της φωνητικής οδού με τη γλωττίδα κλειστή και το $Q(z)$ με τη γλωττίδα ανοιχτή. Έχει αποδειχθεί στο [Soong and Juang, 1984] ότι οι ρίζες αυτών των πολυωνύμων έχουν τις εξής ιδιότητες:

[4] Συνοπτικά, η μερική αυτοσυσχέτιση μεταξύ δύο δειγμάτων $x[n]$, και $x[n+1]$, εκφράζει τη συσχέτιση των δύο κρατώντας τα $x[n+1], \ldots, x[n+i-1]$ σταθερά

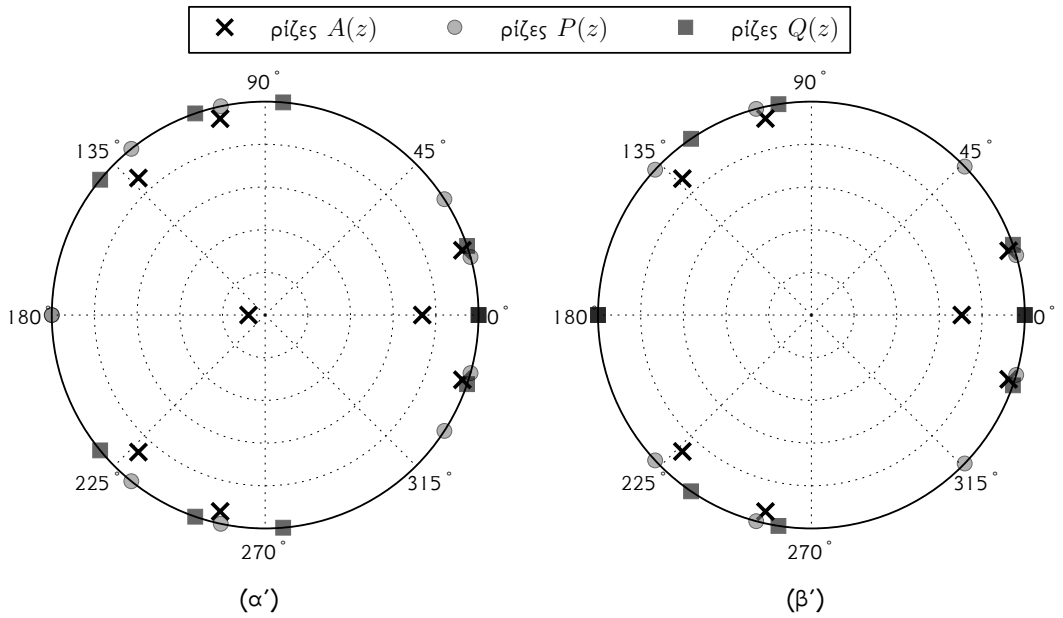
[5] Θυμίζουμε ότι ένα πολυώνυμο $a(x) = \sum_{m=0}^{M} a_m x^m$ λέγεται *παλινδρομικό* εάν $a_m = a_{M-m}$ και *αντιπαλινδρομικό* εάν $a_m = -a_{M-m}$ για $m = 0, 1, \ldots, M$

1. Είναι απλές
2. Προκύπτουν εναλλάξ πάνω στο μοναδιαίο κύκλο

Επίσης, εκμεταλλευόμενοι τις τετριμμένες ρίζες που έχουν στα $z = 1$ ή/και $z = -1$ ορίζουμε τα πολυώνυμα

$$\hat{P}(z) = \frac{P(z)}{1 + z^{-1}} \text{ και } \hat{Q}(z) = \frac{Q(z)}{1 - z^{-1}}, \text{ για } p \text{ άρτιο}$$

$$\hat{Q}(z) = P(z) \quad \text{ και } \hat{Q}(z) = \frac{Q(z)}{1 - z^{-2}}, \text{ για } p \text{ περιττό}$$

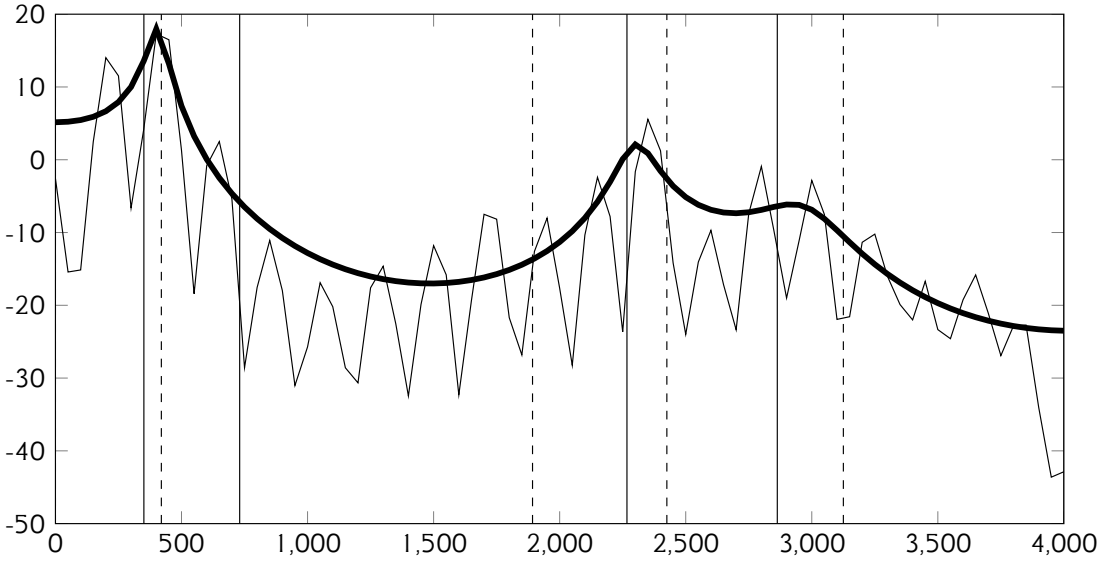


Σχήμα 1.28

Τα πολυώνυμα $\hat{P}(z)$ και $\hat{Q}(z)$ είναι άρτιου βαθμού και συμμετρικά και επειδή οι συντελεστές τους είναι πραγματικοί, οι ρίζες τους είναι συζυγή ζεύγη. Συνεπώς αρκεί να βρούμε τις ρίζες τους στο ένα ημικύκλιο. Οι θέσεις των ριζών $0 < \omega_i < \pi$ αποτελούν τις συχνότητες LSF και στη βιβλιογραφία υπάρχει πλήθος μεθόδων για τον προσδιορισμό τους (π.χ. [Kabal and Ramachandran, 1986]), αφού χρησιμοποιούνται εκτεταμένα για κωδικοποίηση, συμπίεση και μετάδοση

σήματος φωνής. Στο Σχήμα 1.28 βλέπουμε και σχηματικά τον τρόπο με τον οποίο είναι κατανεμημένες οι ρίζες των πολυωνύμων $P(z)$ και $Q(z)$.

Ένα στοιχείο που έχει ιδιαίτερο ενδιαφέρον στο Σχήμα 1.28, είναι ότι όταν κάποια ρίζα του πολυωνύμου πρόβλεψης $A(z)$ έχει πλάτος κοντά στο 1, το "κυκλώνουν" δύο συχνότητες LSF και όσο πιο μεγάλο το πλάτος, τόσο μικρότερη και η απόσταση μεταξύ των δύο αυτών συχνοτήτων. Αυτό ίσως γίνεται πιο ξεκάθαρο στο Σχήμα 1.29 όπου παρατηρούμε ότι κάθε κορυφή κλείνεται από δύο συχνότητες LSF, αλλά και ότι στις κοιλάδες η απόσταση μεγαλώνει.



Σχήμα 1.29: Συχνότητες LSF μαζί με τον αντίστοιχο φασματικό φάκελο από ανάλυση γραμμικής πρόβλεψης τάξης $p = 8$ (Σχήμα 1.28α′)

Σαφματικοί Συνελεστές Γραμμικής Πρόβλεψης Περισσότερα για το "παράξενο" αυτό όνομα θα παρουσιαστούν στην επόμενη ενότητα. Για την ώρα αρκούμαστε στον τρόπο απόκτησης των σαφματικών συντελεστών γραμμικής πρόβλεψης (**L**inear **P**rediction **C**epstral **C**oefficients - LPCC) $\{c_k\}$ από τους συντελεστές πρόβλεψης $\{a_k\}$:

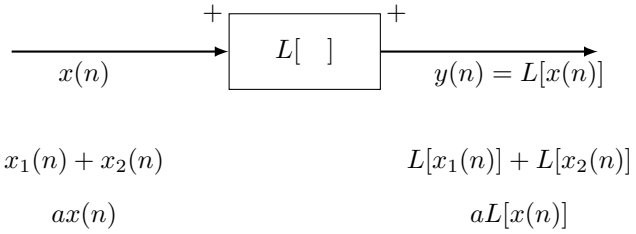
$$c_k = \begin{cases} \ln(G), & k = 0 \\ a_k + \sum_{i=1}^{k-1} \frac{i}{k} c_i a_{k-1}, & 1 \leq k \leq p \end{cases} \tag{1.61}$$

1.6 Ομομορφική Ανάλυση

Σύμφωνα με τα προηγούμενα, κάθε τμήμα ομιλίας *μικρής* χρονικής διάρκειας, μοντελοποιείται αποτελεσματικά σαν να είχε παραχθεί με τη διέγερση ενός γραμμικού και χρονικά αναλλοίωτου συστήματος, είτε από μία σχεδόν-περιοδική ακολουθία παλμών, είτε από ένα τυχαίο σήμα θορύβου. Αφού διέγερση και κρουστική απόκριση ενός τέτοιου συστήματος συνδυάζονται συνελικτικά, μπορούμε να δούμε το πρόβλημα ανάλυσης ομιλίας σαν ένα πρόβλημα διαχωρισμού των συνιστωσών μιας συνέλιξης. Συχνά αποκαλούμε ένα τέτοιο πρόβλημα σαν "αποσυνέλιξη". Στη συνέχεια θα εισάγουμε τις έννοιες των ομομορφικών συστημάτων συνέλιξης και αποσυνέλιξης και θα δούμε εφαρμογές τους στην ανάλυση ομιλίας.

Ομομορφικά συστήματα συνέλιξης και αποσυνέλιξης

Ο όρος "ομομορφικό σύστημα" εμφανίζεται στο [Oppenheim, 1965] όπου αναπτύσσεται μια θεωρία μη-γραμμικών συστημάτων βασισμένη στη θεωρία των γραμμικών διανυσματικών χώρων. Η κεντρική ιδέα αυτής της θεωρίας είναι ότι ορισμένες πράξεις συνδυασμού σημάτων (συγκεκριμένα ο πολλαπλασιασμός και η συνέλιξη) υπακούουν στις ίδιες αρχές όπως η πρόσθεση διανυσμάτων στους γραμμικούς διανυσματικούς χώρους. Από αυτή την παρατήρηση φαίνεται ότι ορισμένες κλάσεις μη-γραμμικών συστημάτων μπορούν να οριστούν στη βάση μιας γενικευμένης αρχής υπέρθεσης. Έτσι, κατ' αναλογία με τη συνήθη αρχή υπέρθεσης για γραμμικά συστήματα (Σχήμα 1.30) όπου έχουμε



Σχήμα 1.30: Αναπαράσταση ενός συστήματος που υπακούει στη συνήθη αρχή υπέρθεσης

$$\begin{aligned} L[x(n)] &= L[x_1(n) + x_2(n)] \\ &= L[x_1(n)] + L[x_2(n)] \\ &= y_1(n) + y_2(n) = y(n) \end{aligned} \tag{1.62α'}$$

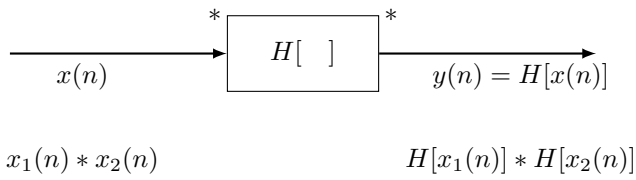
$$L[ax(n)] = aL[x(n)] = ay(n) \tag{1.62β'}$$

με L να είναι ο γραμμικός τελεστής, στα ομομορφικά συστήματα συνέλιξης (Σχήμα 1.31) έχουμε

$$\begin{aligned} H[x(n)] &= H[x_1(n) * x_2(n)] \\ &= H[x_1(n)] * H[x_2(n)] \\ &= y_1(n) * y_2(n) = y(n) \end{aligned} \tag{1.63}$$

Θα μπορούσε να δοθεί και μία εξίσωση παρόμοια με την 1.62β' ώστε να εκφραστεί ο πολλαπλασιαμός με αριθμό με τη γενικευμένη έννοια. Ωστόσο, η ερμηνεία ενός τέτοιου πολλαπλασιασμού δεν είναι αυτονόητη [Schafer, 1968] και καθώς δεν θα τη χρειαστούμε στα παρακάτω, παραλείπεται.

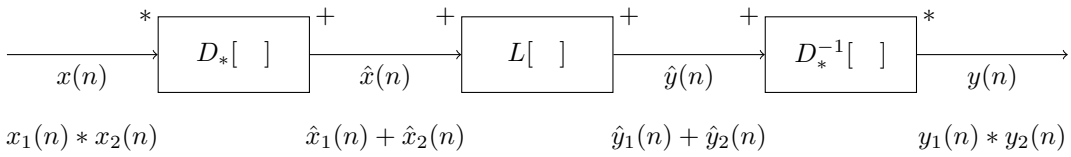
Κάνοντας χρήση τέτοιων συστημάτων, μπορούμε να κατασκευάσουμε *ομομορφικά φίλτρα*, συστήματα δηλαδή που απομακρύνουν τη μη-επιθυμητή συνιστώσα της εισό-



Σχήμα 1.31: Αναπαράσταση ενός ομομορφικού συστήματος συνέλιξης

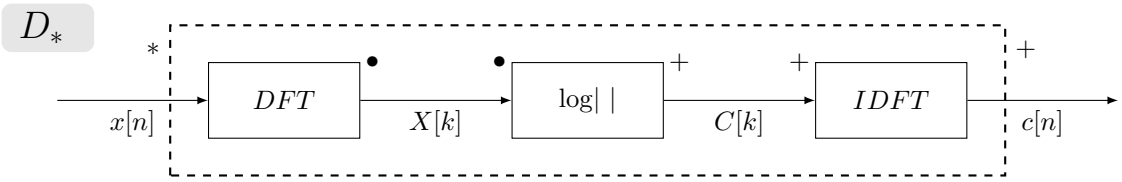
δου. Για παράδειγμα, εάν στην εξίσωση 1.63 η x_1 ήταν η μη-επιθυμητή συνιστώσα θα απαιτούσαμε η συνιστώσα της εξόδου y_1 να είναι ένα μοναδιαίο δείγμα, κάτι εντελώς ανάλογο με την περίπτωση των συνηθισμένων γραμμικών συστημάτων όπως για παράδειγμα στην αντιμετώπιση του πρόβλήματος διαχωρισμού του επιθυμητού σήματος από τον αθροιστικό συνδυασμό σήματος και θορύβου.

Μία σημαντική πτυχή της θεωρίας ομομορφικών συστημάτων, είναι ότι κάθε ομομορφικό σύστημα μπορεί να αναπαρασταθεί ως μια αλυσίδα τριών συστημάτων όπως στο Σχήμα 1.32, όπου απεικονίζεται η *κανονική μορφή* ενός ομομορφικού συστήματος συνέλιξης. Το πρώτο τμήμα της αλυσίδας λαμβάνει εισόδους που συνδυάζονται συνελικτικά και τις μετασχηματίζει σε έναν αθροιστικό συνδυασμό των αντίστοιχων εξόδων. Το δεύτερο τμήμα είναι ένα συμβατικό γραμμικό σύστημα που υπακούει στην αρχή της υπέρθεσης όπως αυτή δίνεται στην Εξ. 1.62 και το τρίτο είναι το αντίστροφο του πρώτου που μετατρέπει τον αθροιστικό συνδυασμό σε συνελικτικό. Η αξία μιας τέτοιας κανονικής μορφής, βρίσκεται στο γεγονός ότι η σχεδίαση του τελικού συστήματος ανάγεται στη σχεδίαση ενός γραμμικού συστήματος, αφού τα *χαρακτηριστικά συστήματα* για ομομορφική συνέλιξη και αποσυνέλιξη ($\boldsymbol{D}_*[\cdot]$ και $\boldsymbol{D}_*^{-1}[\cdot]$ αντίστοιχα), παραμένουν σταθερά. Στην πράξη, για την αποσυνέλιξη των σημάτων διέγερσης και



Σχήμα 1.32: Κανονική μορφή ομομορφικού συστήματος συνέλιξης

κρουστικής απόκρισης του φίλτρου της φωνητικής οδού, χρησιμοποιείται ο Διακριτός Μετασχηματισμός Fourier. Το χαρακτηριστικό σύστημα έχει τη μορφή του Σχήματος 1.33 και το σήμα εξόδου αποκαλείται *σάφμα*. Μεταφράζουμε ελεύθερα σε "σάφμα" τον αγγλικό όρο "cepstrum", έναν αναγραμματισμό της λέξης "spectrum" (φάσμα), που πρωτοεμφανίστηκε στο [Bogert et al., 1963]. Μια ιστορική αναδρομή που περιγράφει τους λόγους που οδήγησαν σε αυτό το περίεργο όνομα και μία σειρά άλλων αντίστοιχων όρων αλλά και τη σχετική θεωρία μπορεί να βρεθεί στο [Oppenheim and Schafer, 2004].

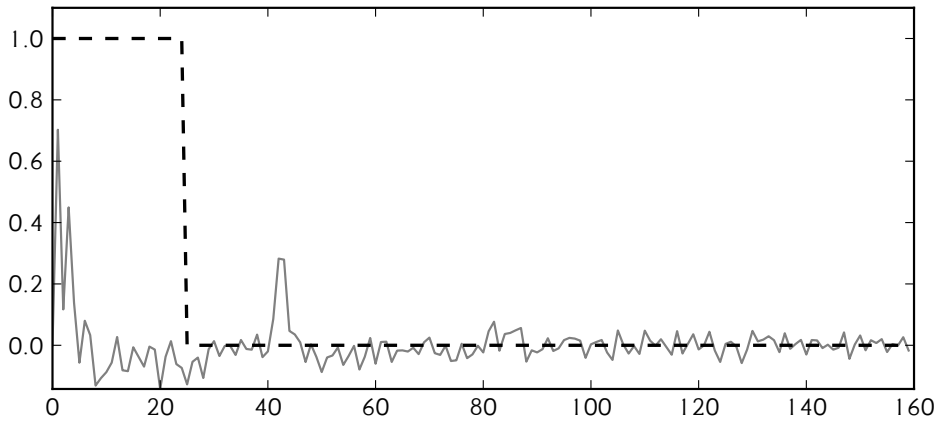


Σχήμα 1.33: Υπολογισμός σάφματος με χρήση Διακριτού Μετασχηματισμού Fourier (DFT)

Επιστρέφουμε στην παρατήρηση ότι η πράξη της συνέλιξης έχει μετατραπεί σε πρόσθεση. Ο διαχωρισμός των δύο τμημάτων γίνεται μέσω ενός απλού "σαφματικού παραθύρου $h[n]$. Για να πάρουμε το τμήμα που αντιστοιχεί στη φωνητική οδό θέτουμε

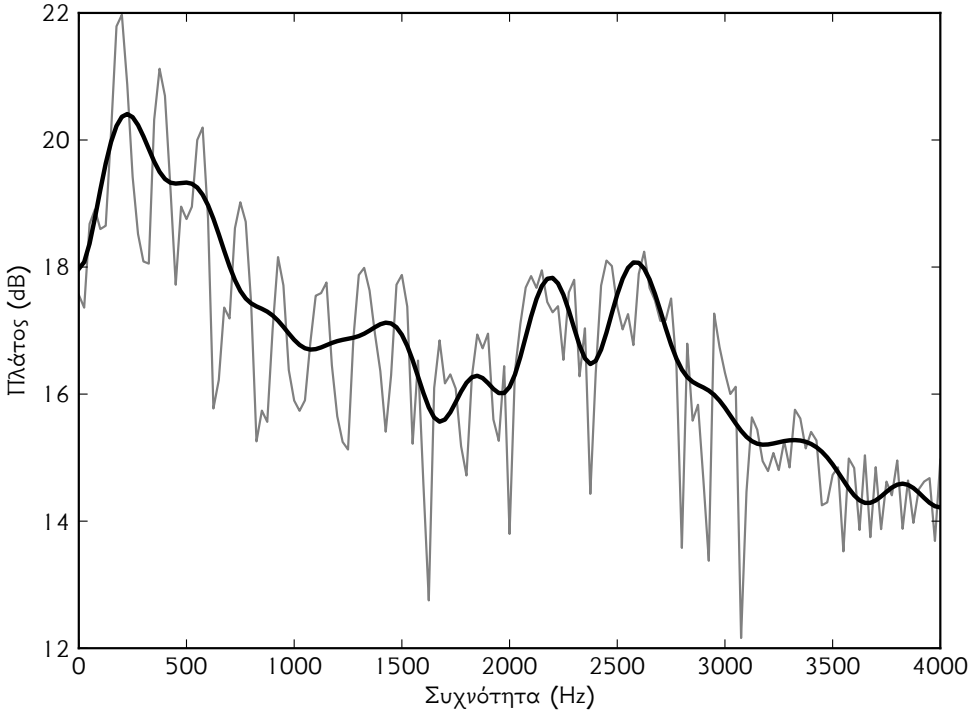
$$l[n] = \begin{cases} 1 & n \leq n_0 \\ 0 & n > n_0 \end{cases} \tag{1.64}$$

για κάποιο n_0 (Σχήμα 1.34). Αυτό οφείλεται στο ότι τα χαρακτηριστικά της φωνητικής οδού αλλάζουν πιο αργά από αυτά του σήματος διέγερσης. Για παράδειγμα, σε περίπτωση έμφωνης ομιλίας το n_0 πρέπει να είναι μικρότερο από την περίοδο του σήματος διέγερσης (τονικό ύψος). Μία από τις δημοφιλείς εφαρμογές της σαφματικής ανάλυσης είναι και η ίδια η εκτίμηση του τονικού ύψους που εντοπίζεται σαν μία κορυφή στο σάφμα (στο Σχήμα 1.34 κοντά στο σημείο $n = 40$). Μέσω του μετασχηματισμού Fourier του αρχικού τμήματος του σάφματος μπορούμε πλέον να



Σχήμα 1.34: Παράθυρο σάφματος

προσεγγίσουμε το φάσμα του πλαισίου ομιλίας με τρόπο αντίστοιχο με αυτόν που ακολουθούμε στην Ανάλυση Γραμμικής Πρόβλεψης (Σχήμα 1.35).



Σχήμα 1.35

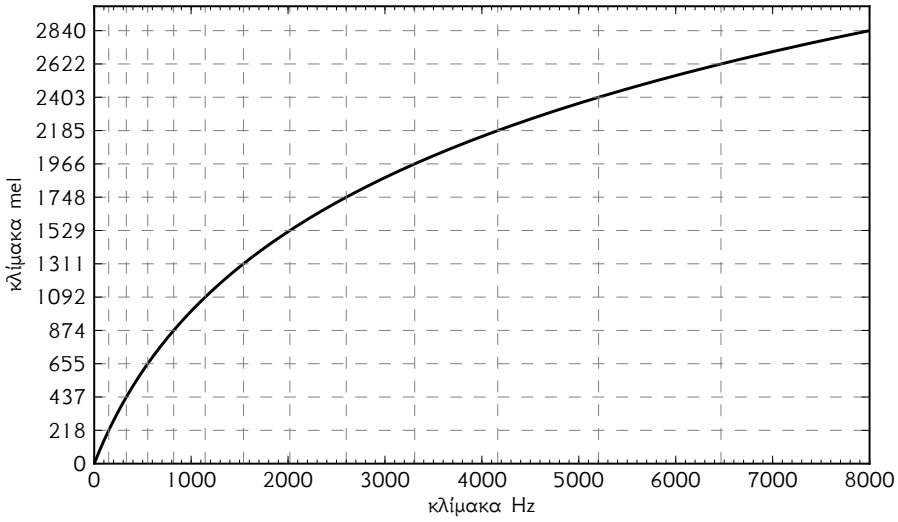
Σαφματικοί Συντελεστές Mel Κλίμακας (MFCC)

Οι σαφματικοί συντελεστές mel κλίμακας (**M**el **Frequency C**epstral **C**oefficients, σε όσα ακολουθούν MFCC), αποτελούν μία από τις πιο δημοφιλείς παραμετροποιήσεις του σήματος ομιλίας. Πρωτοπαρουσιάστηκαν στο [Davis and Mermelstein, 1980] όπου χρησιμοποιήθηκαν στην αναγνώριση μονοσυλλαβικών λέξεων. Έκτοτε έχουν χρησιμοποιηθεί εκτενέστατα και σε άλλες εφαρμογές όπως στην αναγνώριση και διαχωρισμό ομιλητών όπου αποτελούν ένα από τα πιο συνήθη χαρακτηριστικά. Η διαδικασία εξαγωγής των συντελεστών MFCC δοσμένου του μετασχηματισμού DFT μήκους N

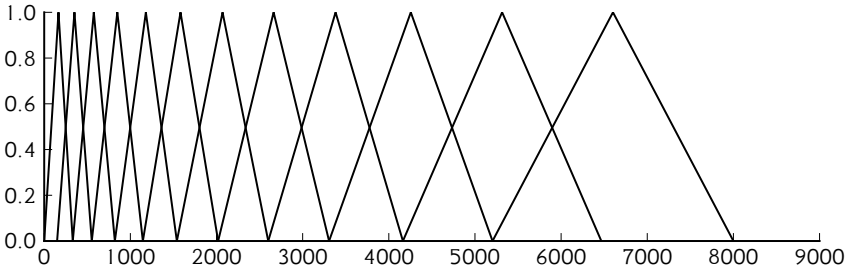
$$X[k] = \sum_{n=0}^{N-1} x[n] \exp\left(-\frac{i2\pi nk}{N}\right), \quad k = 0, 1, \ldots, N-1 \tag{1.65}$$

έχει ως εξής:

1. Το εύρος φάσματος χωρίζεται σε ίσα τμήματα της κλίμακας mel (Σχήμα 1.36) με χρήση M τριγωνικών, συχνά επικαλυπτώμενων φίλτρων (Σχήμα 1.37).



Σχήμα 1.36: Χωρισμός του εύρους συχνοτήτων σε ισομήκη mel τμήματα



Σχήμα 1.37: Τράπεζα 12 mel φίλτρων με επικάλυψη 50%

Καθένα από αυτά τα τριγωνικά φίλτρα ορίζεται από τη σχέση

$$H_i(k) = \begin{cases} 0 & \text{για} \quad k < f_{b_{i-1}} \\ \dfrac{k - f_{b_{i-1}}}{f_{b_i} - f_{b_{i-1}}} & \text{για} \quad f_{b_{i-1}} \leq k \leq f_{b_i} \\ \dfrac{f_{b_{i+1}} - k}{f_{b_{i+1}} - f_{b_i}} & \text{για} \quad f_{b_i} \leq k \leq f_{b_{i+1}} \\ 0 & \text{για} \quad k > f_{b_{i+1}} \end{cases} \quad , i = 1, 2, \ldots, M \tag{1.66}$$

όπου με H_i συμβολίζουμε το i-οστό φίλτρο, με f_{b_i} τα όρια των φίλτρων και με k το k-οστό δείγμα της ακολουθίας DFT.

2. Σαν έξοδος του κάθε φίλτρου ορίζεται η *λογαριθμική ενέργεια* που δίνεται από τη σχέση

$$X_i = \log_{10}\left(\sum_{k=0}^{N-1} |X(k)| \cdot H_i(k)\right) \tag{1.67}$$

3. Τέλος, οι συντελεστές MFCC υπολογίζονται μέσω του Συνημιτονικού Μετασχηματισμού, από τη σχέση

$$C_j = \sum_{i=1}^{M} X_i \cos\left(j(i - 0.5)\frac{\pi}{M}\right) \tag{1.68}$$

ΚΕΦΑΛΑΙΟ 2

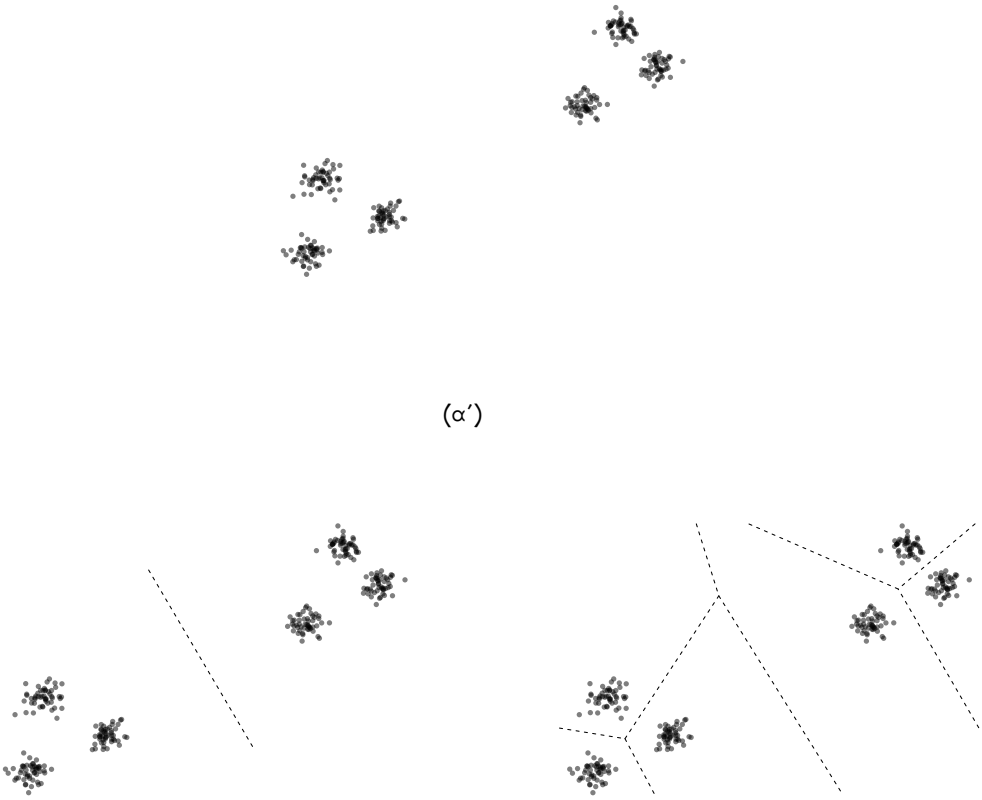
Ομαδοποίηση δεδομένων & Μείωση Διαστάσεων

2.1 Ομαδοποίηση Δεδομένων

Σκοπός όλων των μεθόδων ομαδοποίησης δεδομένων, είναι ο διαχωρισμός ενός συνόλου παρατηρήσεων $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ σε ομάδες, με τρόπο ώστε τα μέλη μιας ομάδας να "μοιάζουν", ενώ τα μέλη διαφορετικών ομάδων να "διαφέρουν". Αυτή είναι μία εργασία την οποία εκτελούμε διαρκώς στην καθημερινότητα, για παράδειγμα χωρίζοντας μία δισκοθήκη ανάλογα με το είδος μουσικής, τακτοποιώντας ρούχα ανάλογα με το είδος τους σε συρτάρια κτλ., φτιάχνοντας έτσι ομάδες "συγγενών" αντικειμένων. Όμως, όταν αυτά τα αντικείμενα αποκτούν τη μορφή σημείων σε έναν χώρο για τα οποία δεν υπάρχει κάποια άλλη πληροφορία, η σημασία των όρων *όμοιο* και *διαφορετικό* δεν είναι πάντα αυτονόητη. Συνήθως, το πόσο διαφορετικά είναι δύο στοιχεία εκφράζεται μέσω μίας απόστασης, συνεπώς στην ομαδοποίηση των δεδομένων $\mathcal{X} \subset \mathbb{R}^d$, χρησιμοποιείται μία κατάλληλη μετρική (π.χ. Ευκλείδια).

Πρέπει να γίνει σαφές ότι η ομαδοποίηση δεδομένων δεν βασίζεται πάνω σε κάποιο αυστηρό θεωρητικό υπόβαθρο και αυτό επειδή δεν υπάρχει κάποιος ορισμός του *τι σημαίνει ομάδα*. Έτσι λοιπόν, κάθε μέθοδος δίνει αυτόν τον ορισμό με το δικό της τρόπο και δημιουργεί ένα σύνολο ομάδων $\mathcal{C}$ ελαχιστοποιώντας κάποια συνάρτηση

κόστους $\boldsymbol{J}(\mathcal{X}, \mathcal{C}\,;\,\boldsymbol{\theta})$. Ακόμα κι έτσι όμως, τα αποτελέσματα της ομαδοποίησης δεν έχουν αξία αν δεν μπορούμε να εξάγουμε κάποιο χρήσιμο συμπέρασμα από αυτά. Για παράδειγμα, τα δεδομένα του Σχήματος 2.1α′ μπορούν να χωριστούν σε δύο (Σχ. 2.1β′) ή έξι (Σχ. 2.1γ′) ομάδες. Το ποια από τις δύο θα προτιμήσουμε εξαρτάται από τη φύση του προβλήματος και από τα συμπεράσματα στα οποία μας οδηγεί καθεμία απο αυτές.



(α′)

(β′) (γ′)

Παρ′ όλα αυτά, αν και φαίνεται ότι ακόμα και σαν ιδέα η ομαδοποίηση έχει κενά, είναι ένα πολύ χρήσιμο εργαλείο σε πολλά επιστημονικά πεδία. Αυτό συμβαίνει επειδή

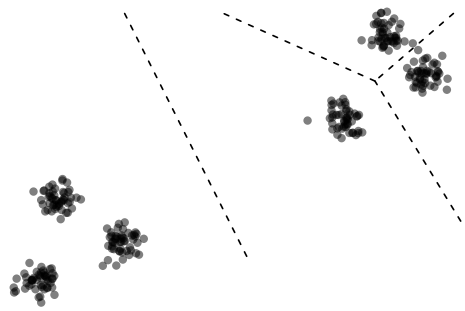
αν και συχνά διαχειριζόμαστε μεγάλο όγκο δεδομένων τα οποία *ξέρουμε* ότι μπορούν να χωριστούν σε ομάδες, πολλές φορές δεν έχουμε τις απαραίτητες πληροφορίες ώστε να κάνουμε τον διαχωρισμό με κάποιο θεωρητικό μοντέλο και έτσι πρέπει να καταφύγουμε σε πιο "εμπειρικές" μεθόδους. Οι εφαρμογές είναι αμέτρητες και χρησιμοποιείται ευρύτατα στην επεξεργασία εικόνων, την ιατρική απεικόνιση, τα χρηματοοικονομικά, τη βιολογία κ.ά. αλλά και στην επεξεργασία ομιλίας (και ειδικότερα στο διαχωρισμό ομιλητών όπως δείχνουμε στο επόμενο κεφάλαιο).

Ο αριθμός των εφαρμογών, αλλά και οι εγγενείς ατέλειες της ομαδοποίησης, έχει οδηγήσει στην ανάπτυξη ενός πολύ μεγάλου αριθμού μεθόδων [Estivill-Castro, 2002]. Η επιλογή μιας μεθόδου ομαδοποίησης δεν είναι πάντα εύκολη, αφού καθεμιά έχει τα πλεονεκτήματα και τα μειονεκτήματά της και εξαρτάται από παράγοντες όπως:

- Το αναμενόμενο σχήμα ή/και ο όγκος των ομάδων, αφού κάποιες μέθοδοι αποδίδουν καλύτερα με ομάδες σφαιρικού σχήματος ή ομάδες με το ίδιο περίπου μέγεθος. Αυτό έχει σχέση με την αντικειμενική συνάρτησή τους και ενώ κάποιες φορές είναι μια ενοχλητική ιδιότητα, κάποιες άλλες βοηθάει στο να αποφευχθεί μια ανομοιόμορφη κατάτμηση που μπορεί να είναι ανεπιθύμητη σύμφωνα με τη φύση του προβλήματος (Σχήμα 2.1).

- Το αν είναι γνωστός εκ των προτέρων ο αριθμός των ομάδων που θα προκύψουν, ή αν πρέπει να εκτιμηθεί. Αυτή είναι μία πολύ σημαντική παράμετρος, αφού στην πλειοψηφία των μεθόδων, γίνεται η υπόθεση ότι ο αριθμός των ομάδων είναι γνωστός. Αυτό δεν ισχύει πάντα και στη βιβλιογραφία υπάρχουν μέθοδοι όπου γίνεται εκτίμηση του αριθμού των ομάδων με διάφορες ευρετικές μεθόδους (π.χ. [Pelleg and Moore, 2000]).

- Το αν το σύνολο των δεδομένων είναι διαθέσιμο, ή ανανεώνεται κατά τη διάρκεια εκτέλεσης της μεθόδου. Αυτό σχετίζεται με την εφαρμογή και συνεπώς έχει να κάνει περισσότερο με την επιλογή αλγορίθμου. Σε κάθε περίπτωση όμως, καθώς η επεξεργασία ροών δεδομένων γίνεται όλο και πιο δημοφιλής,είναι δεδομένη η ανάγκη για τρόπους διαχείρησής τους, αφού αν δεν ληφθεί υπόψη η δυναμική τους φύση, μπορούμε εύκολα να οδηγηθούμε σε απαγορευτικά υπολογιστικά κόστη [O'Callaghan et al., 2001], [Guha et al., 2003].



Σχήμα 2.1: Ομαδοποίηση δεδομένων σε ομάδες διαφορετικού μεγέθους

2.1.1 Η μέθοδος K-μέσων

Μία από τις πιο γνωστές μεθόδους ομαδοποίησης είναι η μέθοδος των K-μέσων. Εδώ, κάθε ομάδα C_i χαρακτηρίζεται από ένα σημείο $\boldsymbol{\mu}_i$ (αναφερόμαστε σε αυτό ως μέσο, κέντρο ή κεντροειδές) και τις παρατηρήσεις που της "ανήκουν".

Δεδομένου ενός συνόλου παρατηρήσεων $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ και ενός ακεραίου $K\,(K \leq n)$, σκοπός της μεθόδου είναι ο διαχωρισμός του $\mathcal{X}$ σε K ξένα υποσύνολα $\mathcal{C} = \{C_1, C_2, \ldots, C_k\}$ ελαχιστοποιώντας τη συνάρτηση κόστους

$$J(\mathcal{X}, \mathcal{C}\,;\, K) = \sum_{i=1}^{K} \sum_{\mathbf{x}_j \in C_i} \left\|\mathbf{x}_j - \boldsymbol{\mu}_i\right\|^2 \tag{2.1}$$

Ένας απλός επαναληπτικός αλγόριθμος για αυτή την ελαχιστοποίηση δίνεται παρακάτω.

Αλγόριθμος 2 Μέθοδος K-μέσων [Lloyd, 1982]

Είσοδος: Το σύνολο των παρατηρήσεων $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^d$
Ο αριθμός των ομάδων $K\,(K \leq N)$

Έξοδος: Σύνολο ομάδων $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$

1: Επιλέγεται τυχαία ένα αρχικό σύνολο κεντροειδών

$$\{\boldsymbol{\mu}_1^{(0)}, \boldsymbol{\mu}_2^{(0)}, \ldots, \boldsymbol{\mu}_K^{(0)}\} \subseteq \mathcal{X}$$

2: Δημιουργούνται οι ομάδες σύμφωνα με τη σχέση

$$C_i^{(t)} = \left\{\mathbf{x}_p \,:\, \left\|\mathbf{x}_p - \boldsymbol{\mu}_i^{(t)}\right\| \leq \left\|\mathbf{x}_p - \boldsymbol{\mu}_j^{(t)}\right\| \,;\, \forall 1 \leq j \leq K\right\}$$

3: Ανανεώνονται οι θέσεις των κεντροειδών σύμφωνα με τη σχέση

$$\boldsymbol{\mu}_i^{(t+1)} = \frac{1}{|C_i^{(t)}|} \sum_{\mathbf{x}_j \in C_i^{(t)}} \mathbf{x}_j$$

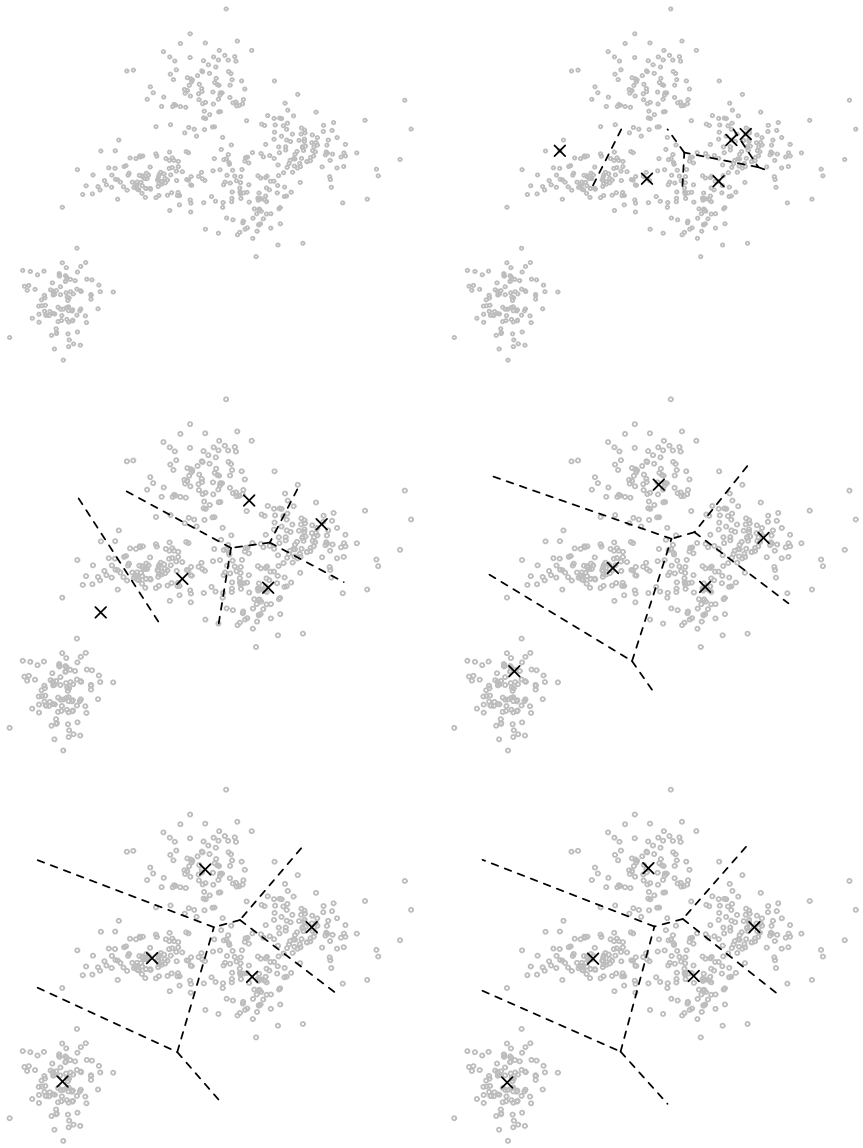
4: Επαναλαμβάνονται τα βήματα 2 και 3 μέχρι να μην υπάρχουν αλλαγές στο σύνολο $\mathcal{C}$

Αν και όπως είπαμε η μέθοδος των K-μέσων είναι από τις πιο επιδραστικές μεθόδους στην ομαδοποίηση δεδομένων, σπάνια εφαρμόζεται με τον τρόπο που περιγράφεται στον Αλγ. 2. Υπάρχουν δύο λόγοι για αυτό:

1. Η αρχικοποίηση των κέντρων, όντας τυχαία, μπορεί εύκολα να "παγιδεύσει" τη μέθοδο σε τοπικό ελάχιστο της αντικειμενικής της συνάρτησης, δίνοντας ένα μη-επιθυμητό διαχωρισμό. Οι λύσεις που έχουν προταθεί στη βιβλιογραφία, περιλαμβάνουν μεταξύ άλλων διαφορετικό τρόπο αρχικοποίησης ή εκτέλεση πολλών επαναλήψεων του αλγορίθμου.

2. Σε κάθε επανάληψη του αλγορίθμου, απαιτείται ο υπολογισμός όλων των αποστάσεων μεταξύ παρατηρήσεων και κέντρων. Αυτό έχει σαν αποτέλεσμα ο χρόνος εκτέλεσής του να αυξάνει εκθετικά όσο μεγαλώνει ο αριθμός των παρατηρήσεων n και η διάστασή τους d.

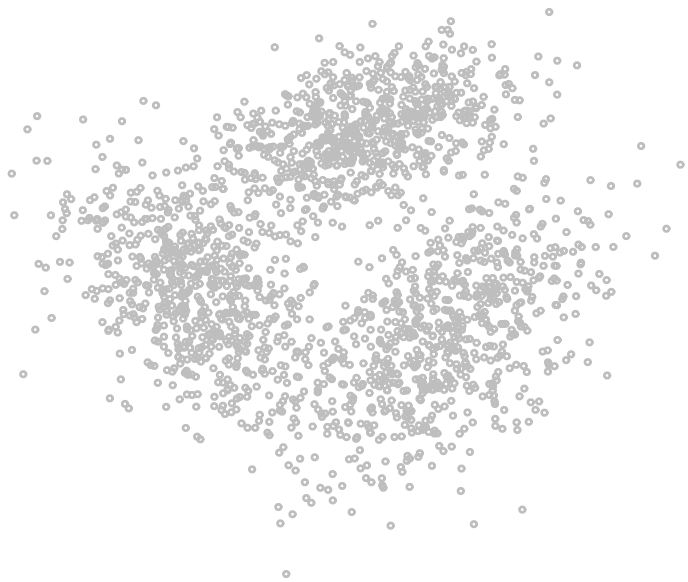
Για να αντιμετωπιστούν αυτά τα προβλήματα έχουν κατά καιρούς προταθεί πολλές παραλλαγές αυτού του αλγορίθμου (π.χ. [Elkan, 2003], [Arthur and Vassilvitskii, 2007], [Hamerly, 2010], [Bahmani et al., 2012]), προτείνοντας πιο αποδοτικές υλοποιήσεις. Η μέθοδος των K-μέσων παραμένει διαχρονικά δημοφιλής και χρησιμοποιείται ευρύτατα είτε μόνη της, είτε σαν αρχικό στάδιο κάποιας πιο σύνθετης μεθόδου.



Σχήμα 2.2: Επαναλήψεις του αλγόριθμου K-μέσων ($K = 5$)

2.1.2 Μέθοδος Ασαφών K-μέσων

Οι μέθοδοι ασαφούς ομαδοποίησης χρησιμοποιούν τις ιδέες των ασαφών συνόλων [Zadeh, 1965], όπου ένα αντικείμενο μπορεί να ανήκει *μερικώς* σε ένα σύνολο. Αυτό σημαίνει ότι επιτρέπεται στο κάθε στοιχείο του $\mathcal{X}$ να ανήκει σε παραπάνω από μία ομάδες με διαφορετικό βαθμό και ακριβώς εκεί δικαιολογείται ο όρος "ασαφές", αφού δεν είναι πλέον σαφές το που ανήκει κάθε δείγμα και τα όρια των ομάδων δεν είναι ευδιάκριτα. Οι ασαφείς μέθοδοι χρησιμοποιούνται συχνά σε εφαρμογές, αφού αρκετές φορές μία ασαφής κατάτμηση ερμηνεύεται με πιο φυσικό τρόπο λόγω της παρουσίας ενός παράγοντα αβεβαιότητας που ενώ όπως έχουμε πει είναι ούτως ή άλλως μέρος της διαδικασίας ομαδοποίησης, δεν υπάρχει σε "σκληρές" μεθόδους (όπως αυτή των K-μέσων). Ας δούμε για παράδειγμα τα δεδομένα του Σχήματος 2.3. Μπορούμε να καταλάβουμε ότι υπάρχουν τρεις ομάδες, αλλά τα όριά τους δεν είναι ευδιάκριτα.



Σχήμα 2.3

Οι διαφορετικές προσεγγίσεις μπορούν να γίνουν ορατές δίνοντας έναν από έναν ορισμό για τη σκληρή και την ασαφή ομαδοποίηση. Χρησιμοποιώντας συμβολισμό από την κλασσική θεωρία συνόλων, μία σκληρή ομαδοποίηση του $\mathcal{X}$ είναι μια οικογένεια

υποσυνόλων $\{A_i | 1 \leq i \leq K\} \subset \mathcal{P}(\mathcal{X})$ με τις εξής ιδιότητες:

$$\bigcup_{i=1}^{K} A_i = \mathcal{X} \tag{2.2α'}$$

$$A_i \cap A_j = \emptyset, \quad 1 \leq i \neq j \leq K \tag{2.2β'}$$

$$\emptyset \subset A_i \subset \mathcal{X}, \quad 1 \leq i \leq K \tag{2.2γ'}$$

Το αποτέλεσμα της ομαδοποίησης είναι ένας $K \times N$ *πίνακας συμμετοχής* $U = [u_{ij}]$ με τα στοιχεία του να δηλώνουν τη συμμετοχή της κάθε παρατήρησης σε κάθε ομάδα. Πιο συγκεκριμένα τα στοιχεία του u_{ij} ικανοποιούν τις συνθήκες

$$u_{ij} \in \{0, 1\}, \quad 1 \leq i \leq K, 1 \leq j \leq N \tag{2.3α'}$$

$$\sum_{i=1}^{K} u_{ij} = 1, \quad 1 \leq j \leq N \tag{2.3β'}$$

$$0 < \sum_{j=1}^{N} u_{ik} < N, \quad 1 \leq i \leq K \tag{2.3γ'}$$

Συγκεντρώνοντας τα παραπάνω, ορίζεται ο χώρος πινάκων σκληρής ομαδοποίησης

$$M_{hc} = \left\{ U \in \mathbb{R}^{K \times N} \middle| u_{ij} \in 0, 1, \forall i, k; \sum_{i=1}^{K} u_{ij} = 1, \forall j; 0 < \sum_{j=1}^{N} u_{ij} < N, \forall i \right\} \tag{2.4}$$

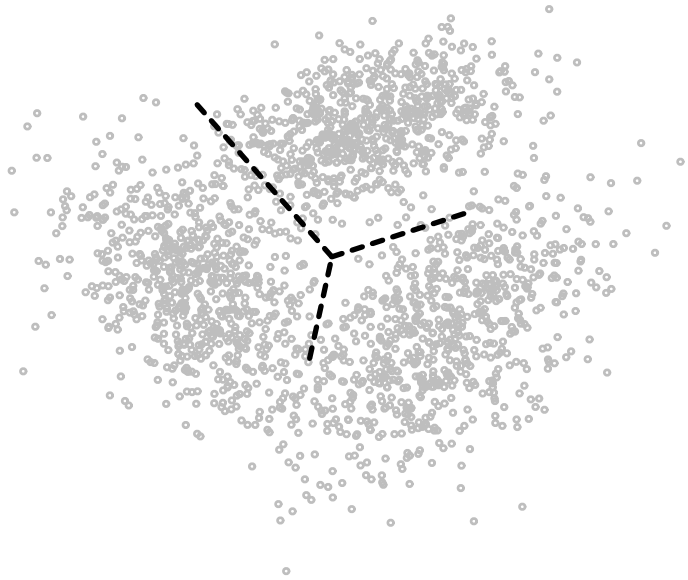
Μία σκληρή ομαδοποίηση των δεδομένων του Σχήματος 2.3 δίνεται στο Σχήμα 2.4. Εδώ φαίνεται και η αδυναμία των αυστηρών ορίων των ομάδων, μεγάλος αριθμός παρατηρήσεων βρίσκεται ακριβώς ή πολύ κοντά σε αυτά, γεγονός που μειώνει την "εμπιστοσύνη" που μπορούμε να έχουμε σε μία σκληρή ομαδοποίηση, παρά το γεγονός ότι σχηματίζονται οι τρεις ομάδες που διαισθητικά είχαμε εντοπίσει στην αρχή.

Οι διαφορές στην προσέγγιση της ασαφούς ομαδοποίησης εντοπίζονται στις εξισώσεις (2.2β') και (2.3α'). Η πρώτη προφανώς δεν ισχύει, ενώ η δεύτερη γίνεται

$$u_{ij} \in [0, 1], \quad 1 \leq i \leq K, 1 \leq j \leq N \tag{2.5}$$

δηλώνοντας έτσι τη *μερική συμμετοχή* μιας παρατήρησης σε μια ομάδα. Σε αντιστοιχία με την εξίσωση (2.4) ορίζεται ο χώρος πινάκων συμμετοχής ασαφούς ομαδοποίησης

$$M_{fc} = \left\{ U \in \mathbb{R}^{K \times N} \middle| u_{ij} \in [0, 1], \forall i, k; \sum_{i=1}^{K} u_{ij} = 1, \forall j; 0 < \sum_{j=1}^{N} u_{ij} < N, \forall i \right\} \tag{2.6}$$



Σχήμα 2.4

Οι ομάδες σε αυτή την περίπτωση έχουν τη μορφή του Σχήματος 2.5, όπου χρησιμοποιούμε την ένταση του χρώματος για να αναπαραστήσουμε το ποσοστό συμμετοχής.

Η μέθοδος που χρησιμοποιήσαμε για αυτή την ομαδοποίηση λέγεται *μέθοδος ασαφών K-μέσων* και μπορεί να θεωρηθεί μία επέκταση της σκληρής μεθόδου. Εδώ η συνάρτηση κόστους που ελαχιστοποιείται είναι η

$$J(\mathcal{X}, U, V\ ;\ K, \mathrm{A}, m) = \sum_{j=1}^{K} \sum_{i=1}^{N} (u_{ji})^m \|\mathbf{x}_i - \mathbf{v}_j\|_A^2 \tag{2.7}$$

όπου

$$U = \text{ασαφής ομαδοποίηση του } \mathcal{X}\text{:}\ \ U \in M_{fc} \tag{2.8α'}$$

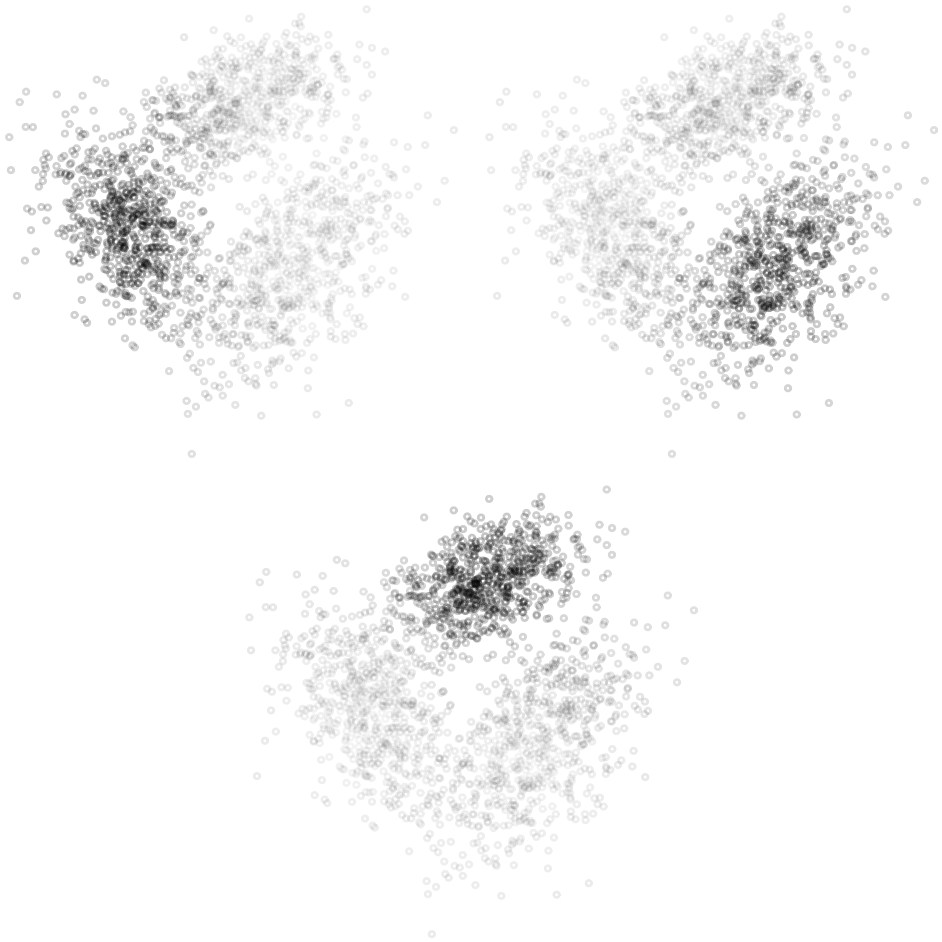
$$K = \text{αριθμός ομάδων στο } \mathcal{X}\text{:}\ \ 2 \leq K \leq N \tag{2.8β'}$$

$$m = \text{εκθέτης ασάφειας:}\ \ 1 \leq m < \infty \tag{2.8γ'}$$

$$V = \left\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K\right\} = \text{ σύνολο κέντρων ομάδων} \tag{2.8δ'}$$

$$\mathrm{A} = \text{θετικά ορισμένος } d \times d \text{ πίνακας σχήματος} \tag{2.8ε'}$$

$$\|\cdot\|_A = \text{επαγόμενη νόρμα του πίνακα } A \tag{2.8στ'}$$



Σχήμα 2.5

Το πρώτο από τα δύο καινούργια στοιχεία σε σχέση με τη μέθοδο K-μέσων, είναι ο πίνακας συντελεστών βάρους A ο οποίος ελέγχει το βέλτιστο σχήμα των ομάδων που έχουμε υποθέσει. Η επαγόμενη από αυτόν A-νόρμα ορίζεται από το εσωτερικό γινόμενο

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \mathbf{x}^T A \mathbf{y} \tag{2.9}$$

δηλαδή

$$\left\| \mathbf{x}_i - \mathbf{v}_j \right\|_A^2 = (\mathbf{x}_i - \mathbf{v}_j)^T A (\mathbf{x}_i - \mathbf{v}_j) \tag{2.10}$$

Φυσικά, υπάρχουν άπειρες επιλογές για τον πίνακα A, αλλά στην πράξη μόνο

λίγες χρησιμοποιούνται. Ενδεικτικά, έστω

$$\mathbf{m} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \tag{2.11α'}$$

$$S_x = \sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T \tag{2.11β'}$$

να είναι ο δειγματικός μέσος και δειγματικός πίνακας συνδιακύμανσης του $\mathcal{X}$, $\{a_j\}$ οι ιδιοτιμές του S_x, D_x ο διαγώνιος πίνακας με στοιχεία $(d_x)_{jj} = a_{jj}$ και I ο μοναδιαίος πίνακας. Οι τρεις πιο δημοφιλείς επιλογές του πίνακα A είναι οι $A = I$, $A = D_x^{-1}$ και $A = S_x^{-1}$. Μία εκτενής περιγραφή των ιδιοτήτων της κάθε επιλογής γίνεται στο [Bezdek, 1981].

Η άλλη νέα παράμετρος είναι ο συντελεστής ασάφειας m. Μπορεί να δειχθεί ότι για $m \to 1$, η ομαδοποίηση είναι σκληρή, ενώ αντίθετα για $m \to \infty$ έχουμε $u_{ij} \to 1/K\ \forall i, j$. Συνήθως οι επιλογές που γίνονται στην πράξη περιορίζονται στο διάστημα $1.5 \leq m \leq 3.0$.

Ο αλγόριθμος που δίνουμε παρακάτω αποτελεί ίσως την πιο δημοφιλή επιλογή για την ελαχιστοποίηση της συνάρτησης κόστους (2.7)

Αλγόριθμος 3 Μέθοδος ασαφών K-μέσων [Bezdek et al., 1984]

Είσοδος: Το σύνολο των παρατηρήσεων $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^d$
Ο αριθμός των ομάδων $K\,(2 \leq K \leq N)$
Ο πίνακας σχήματος A
Η παράμετρος ασάφειας m

Έξοδος: Ο πίνακας συμμετοχής U

1: Αρχικοποίηση πίνακα συμμετοχής

$$t = 0: \quad \hat{U}^{(0)} \in M_{fc}$$

2: Υπολογισμός κέντρων $\hat{\mathbf{v}}_j^{(t)}, \quad j = 1, 2, \ldots, K$

$$\hat{\mathbf{v}}_j = \frac{\sum_{i=1}^{N} \left(\hat{u}_{ji}^{(t)}\right)^m \mathbf{x}_i}{\sum_{i=1}^{N} \left(\hat{u}_{ji}^{(t)}\right)^m}, \quad j = 1, \ldots, K$$

3: Υπολογισμός πίνακα $\hat{U}^{(t+1)} = \left[\hat{u}_{ji}^{(t+1)}\right]$

$$\hat{u}_{ji}^{(t+1)} = \left[\sum_{l=1}^{K} \left(\frac{\hat{d}_{ji}}{\hat{d}_{li}}\right)^{2/(m-1)}\right]^{-1}, \quad i = 1, \ldots, N;\ j = 1, \ldots, K$$

με

$$d_{ji}^2 = \left\|\mathbf{x}_i - \mathbf{v}_j^{(t)}\right\|_A^2 = (\mathbf{x}_i - \mathbf{v}_j)^T A (\mathbf{x}_i - \mathbf{v}_j^{(t)})$$

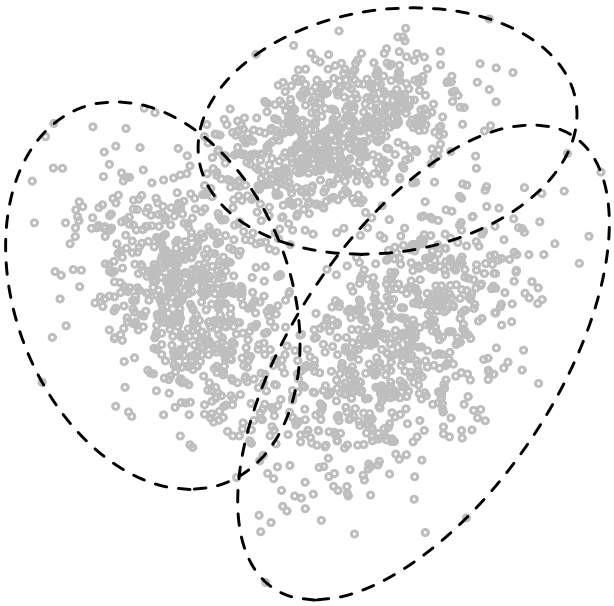
4: Επαναλαμβάνονται τα βήματα 2 και 3 (με $t = t + 1$) έως ότου

$$\|\hat{U}^{(t+1)} - \hat{U}^{(t)}\| < \epsilon$$

για κάποια νόρμα πινάκων και $\epsilon > 0$.

5: $U = \hat{U}^{(t+1)}$

Η ειρωνία με τις μεθόδους ασαφούς ομαδοποίησης, είναι ότι αν και περιγράφουν με πιο φυσικό τρόπο τα δεδομένα, τις περισσότερες φορές *χρειαζόμαστε* μία σκληρή ομαδοποίηση, με την έννοια ότι πρέπει να πάρουμε κάποια απόφαση για το που ανήκει η κάθε παρατήρηση. Ας υποθέσουμε για παράδειγμα, ότι οι παρατηρήσεις ήταν χαρακτηριστικά από ένα σήμα ομιλίας και οι ομάδες αντιστοιχούν στους ομιλητές που συμμετέχουν. Η ομαδοποίηση θα είχε σκοπό να αντιστοιχίσει κάθε παρατήρηση σε κάποιον ομιλητή και δεν έχει νόημα να πούμε ότι κάποια παρατήρηση προέρχεται μερικώς από κάποιον και αυτό που γίνεται συνήθως σε τέτοιες περιπτώσεις, είναι να αντιστοιχούμε την κάθε παρατήρηση στην ομάδα όπου συμμετέχει περισσότερο. Αλλά ακόμα και έτσι, η ασαφής ομαδοποίηση δεν είναι "χαμένος κόπος", αφού με τον τρόπο που υπολογίζονται τα ποσοστά συμμετοχής, επιτρέπεται η αλληλοεπικάλυψη των ομάδων, δίνοντας έτσι αποτελέσματα που δεν θα μπορούσαμε μέσω μεθόδων όπως των K-μέσων (Σχήμα 2.6).



Σχήμα 2.6

2.1.3 Επαλήθευση Ομαδοποίησης

Αναφέραμε και πιο πριν ότι όταν χρησιμοποιούνται επαναληπτικές μέθοδοι όπως αυτές που παρουσιάσαμε, η αναζήτηση του ελαχίστου της αντικειμενικής συνάρτησης εξαρτάται σε μεγάλο βαθμό από την αρχικοποίηση των παραμέτρων. Όταν οι διαστάσεις είναι λίγες μπορούμε να κρίνουμε με απλό οπτικό έλεγχο το κατά πόσο τα αποτελέσματα είναι ικανοποιητικά, αλλά κάτι τέτοιο συνήθως δεν είναι επιθυμητό ή είναι και αδύνατο. Για αυτό το λόγο έχουν προταθεί διάφορα κριτήρια που ποσοτικοποιούν την "ποιότητα" της ομαδοποίησης [Vendramin et al., 2010], ώστε αυτός ο έλεγχος να γίνεται με ελάχιστο κόπο ή και αυτόματα. Κάποια χαρακτηριστικά παραδείγματα είναι τα εξής:

Δείκτης Dunn Ο δείκτης Dunn είναι ένα από τα πιο απλά μέτρα επαλήθευσης ομαδοποίησης. Ορίζεται ως

$$DN = \min_{p,q \in \{1,\ldots,K\}} \left\{ \frac{\delta_{p,q}}{\max_{l \in \{1,\ldots,K\}} \Delta_l} \right\} \tag{2.12}$$

όπου Δ_l η διάμετρος της l-οστής ομάδας και $\delta_{p,q}$ η απόσταση των ομάδων p, q (η ελάχιστη απόσταση μεταξύ δύο στοιχείων από τις ομάδες p και q). Είναι

σαφές ότι μία μεγάλη τιμή του δείκτη Dunn επιτυγχάνεται για συμπαγείς ομάδες (μικρός παρονομαστής) με μεγάλη απόσταση μεταξύ τους (μεγάλος αριθμητής) και υποδεικνύει μια επιτυχή ομαδοποίηση.

Πλάτος Σιλουέτας Θεωρούμε ότι το j-οστό στοιχείο $\mathbf{x}_j$ του $\mathcal{X}$ ανήκει στην p-οστή ομάδα. Συμβολίζουμε με $a_{p,j}$ τη μέση απόστασή του από τα υπόλοιπα στοιχεία της ίδιας ομάδας και με $d_{q,j}$ τη μέση απόσταση από τα στοιχεία μιας ομάδας $q \neq p$. Τέλος με $b_{p,j}$ συμβολίζουμε το ελάχιστη τιμή $d_{q,j}$ για $q = 1, \ldots, K;\ \ q \neq p$. Η *σιλουέτα* του στοιχείου $\mathbf{x}_j$ ορίζεται ως

$$s_{\mathbf{x}_j} = \frac{b_{p,j} - a_{p,j}}{\max\left\{a_{p,j}, b_{p,j}\right\}} \tag{2.13}$$

με τον παρονομαστή να λειτουργεί σαν παράγοντας κανονικοποίησης. Το πλάτος σιλουέτας ορίζεται ως ο μέσος όρος των $s_{\mathbf{x}_j}$, δηλαδή

$$SWC = \frac{1}{N}\sum_{j=1}^{N} s_{\mathbf{x}_j} \tag{2.14}$$

Και σε αυτή την περίπτωση, μία "σωστή" ομαδοποίηση οδηγεί σε υψηλή τιμή πλάτους σιλουέτας.

Σημείωση 2. *Στην περίπτωση που ένα στοιχείο $\mathbf{x}_j$ αποτελεί από μόνο του μία ομάδα, έχουμε εξ' ορισμού $s_{\mathbf{x}_j}$ για να αποφευχθεί να θεωρηθεί ως βέλτιστη η ομαδοποίηση όπου κάθε στοιχείο είναι και ομάδα $k = N$.*

Δείκτης Xie-Beni Ο δείκτης Xie-Beni αφορά την ασαφή ομαδοποίηση (αν και τα υπόλοιπα μέτρα μπορούν να χρησιμοποιηθούν επίσης) και ορίζεται ως

$$V_{XB} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{K} u_{ji}\left\|\mathbf{x}_i - \mathbf{v}_j\right\|^2}{N \cdot \min_{i\neq j}\left\|\mathbf{v}_i - \mathbf{v}_j\right\|^2} \tag{2.15}$$

υιοθετώντας το συμβολισμό που χρησιμοποιήσαμε στην προηγούμενη ενότητα (εδώ βλέπουμε ότι μικρή τιμή του δείκτη μεταφράζεται σαν επιτυχής ομαδοποίηση). Αν και ο δείκτης Xie-Beni μπορεί να χρησιμοποιηθεί άμεσα σε μία ασαφή

ομαδοποίηση, κληρονομεί ταυτόχρονα και τις αδυναμίες της. Πιο συγκεκριμένα, για $m \to \infty$ έχουμε

$$\lim_{m\to\infty} \mathbf{v}_i = \frac{\sum_{i=1}^{N} \mathbf{x}_i}{N} = \bar{\mathbf{v}} \Rightarrow \lim_{m\to\infty} V_{XB} = \infty \tag{2.16}$$

και η συμπεριφορά του δείκτη γίνεται απρόβλεπτη. Επίσης, η έλλειψη ενός παράγοντα ποινής στον αριθμό των ομάδων ($\lim_{K\to N} V_{XB} = 0$) έχει οδηγήσει στη δημιουργία επεκτάσεών του (π.χ [Tang et al., 2005]), αλλά σε κάθε περίπτωση παραμένει ένα δημοφιλές κριτήριο.

2.2 Μείωση Διαστάσεων

Στη βιβλιογραφία συχνά απαντάται ο όρος *κατάρα της διαστατικότητας (curse of dimensionality)*. Ο όρος αποδίδεται στον Richard Bellman [Bellman, 1961], που περιγράφει με αυτόν τον τρόπο τα προβλήματα που σχετίζονται με την ανάλυση δεδομένων σε χώρους μεγάλων διαστάσεων και την αδυναμία ασφαλούς εκτίμησης στατιστικών παραμέτρων, ειδικά όταν ο αριθμός των δεδομένων είναι σχετικά μικρός. Η "κατάρα" όμως δεν αφορά μόνο τη στατιστική εκτίμηση. Για παράδειγμα, ένα πρόβλημα που αφορά την ομαδοποίηση δεδομένων, είναι πως η έννοια της απόστασης χάνει το νόημά της καθώς $\lim_{d\to\infty}(d_{max} - d_{min})/d_{min} \to 0$ και η διάκριση του κοντινού από το μακρινό σημείο γίνεται ανούσια. Είναι λοιπόν φυσικό να επιθυμούμε να εργαζόμαστε με μικρό αριθμό διαστάσεων, αφού εκτός των άλλων, είναι πιο εύκολο να σκεφτούμε σε λίγες διαστάσεις αλλά και να οπτικοποιήσουμε τα δεδομένα. Δυστυχώς πολλές φορές είναι αδύνατο να αποφύγουμε τις πολλές διαστάσεις, είτε επειδή η φύση του προβλήματος που μελετάμε επιβάλλει πολλές μεταβλητές, είτε επειδή μη γνωρίζοντας ακριβώς τις μεταβλητές που εμπλέκονται, το μοντελοποιούμε χρησιμοποιώντας πολλές μεταβλητές ελπίζοντας πως θα φανούν χρήσιμες. Αν με τα προηγούμενα συνυπολογίσουμε και τις παράπλευρες συνέπειες που έχει η χρήση δεδομένων μεγάλης διάστασης, όπως η εκθετική συχνά αύξηση του υπολογιστικού κόστους της ομαδοποίησης ή το κόστος αποθήκευσης, φαίνεται λογικό ότι το θέμα της διαστατικότητας συχνά απασχολεί την επιστημονική κοινότητα, κάτι που έχει σαν αποτέλεσμα την ανάπτυξη μεθόδων για τη μείωση διαστάσεων σε προβλήματα ομαδοποίησης.

2.2.1 Ανάλυση Κύριων Συνιστωσών

Η Ανάλυση Κύριων Συνιστωσών (**P**rincipal **C**omponents **A**nalysis - PCA) είναι πιθανόν η πιο δημοφιλής τεχνική πολυμεταβλητής στατιστικής ανάλυσης και χρησιμοποιείται κατά κόρον σε όλα τα επιστημονικά πεδία, σχεδόν σε κάθε εφαρμογή που περιλαμβάνει επεξεργασία δεδομένων. Ταυτόχρονα είναι και από τις πιο παλιές: Οι ιδέες που εφαρμόζονται θεωρείται ότι πρωτοπαρουσιάστηκαν από τον Pearson το 1901 [Pearson, 1901] και η θεωρία που χρησιμοποιείται (φασματική ανάλυση πίνακα) έχει θεμελιωθεί πολύ πιο πριν. Πάντως, ο όρος "κύρια συνιστώσα" αποδίδεται στον Hotelling [Hotelling, 1933] ο οποίος εισήγαγε και την πιο σύγχρονη εφαρμογή αυτής της τεχνικής.

Ας υποθέσουμε ότι έχουμε ένα σύνολο δεδομένων

$$\mathcal{X} = \{x_1, x_2, \dots, x_N\}, \quad x_i \in \mathbb{R}^d, i = 1, 2, \dots, N \tag{2.17}$$

εκφρασμένα ως προς μια βάση του $\mathbb{R}^d$, $\{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_d\}$. Η ανάλυση κύριων συνιστωσών στην πραγματικότητα δεν είναι μέθοδος μείωσης διαστάσεων, αλλά μία μέθοδος προσδιορισμού μιας εναλλακτικής, "καταλληλότερης" βάσης για τα δεδομένα μας. Η βασική παραδοχή που γίνεται, είναι ότι οι κατευθύνσεις στον $\mathbb{R}^d$ όπου τα δεδομένα παρουσιάζουν μεγάλη διασπορά φέρουν και την περισσότερη πληροφορία και αντίθετα, οι κατευθύνσεις όπου παρουσιάζεται μικρή διασπορά είναι λιγότερο σημαντικές. Κατά συνέπεια οι τελευταίες μπορούν να αγνοηθούν, οδηγώντας σε μείωση διαστάσεων με μικρές απώλειες πληροφορίας. Σκοπός της ανάλυσης είναι ο προσδιορισμός όλων αυτών των κατευθύνσεων και η ιεράρχησή τους ως προς τη "σημασία" τους.

Χωρίς βλάβη της γενικότητας θα θεωρήσουμε ότι για την τυχαία παρατήρηση $\mathbf{x}$ ισχύει $\mathcal{E}[\mathbf{x}] = \mathbf{0}$ (εναλλακτικά χρησιμοποιούμε $\tilde{\mathbf{x}} = \mathbf{x} - \mathcal{E}[\mathbf{x}]$). Αφού σημειώσουμε ότι η προβολή ενός δείγματος $\mathbf{x}$ στη διεύθυνση του διανύσματος $\mathbf{w}$ είναι $\mathbf{y} = \mathbf{w}^T\mathbf{x}$, η πρώτη κύρια συνιστώσα $\mathbf{w}_1$ ορίζεται ως η κατεύθυνση $\mathbf{w}$ όπου μεγιστοποιείται η διασπορά του $\mathbf{y} = \mathbf{w}^T\mathbf{x}$, δηλαδή

$$\mathbf{w}_1 = \underset{\|\mathbf{w}\|=1}{\operatorname{argmax}}\, J(\mathbf{w}) \tag{2.18}$$

με

$$\begin{aligned} J(\mathbf{w}) &= \mathrm{Var}\left[\mathbf{w}^T\mathbf{x}\right] \\ &= \mathcal{E}\left[\left(\mathbf{w}^T\mathbf{x}\right)^2\right] \\ &= \mathcal{E}\left[\mathbf{w}^T\mathbf{x}\mathbf{x}^T\mathbf{w}\right] \\ &= \mathbf{w}^T\,\mathcal{E}\left[\mathbf{x}\mathbf{x}^T\right]\mathbf{w} \\ &= \mathbf{w}^T S\mathbf{w} \end{aligned} \tag{2.19}$$

Σημείωση 3. *Ο περιορισμός του μέτρου* $\|\mathbf{w}\|$ *στην παραπάνω εξίσωση είναι απαραίτητος, αφού σε αντίθετη περίπτωση έχουμε* $J(\mathbf{w}) \to \infty$ *για* $\|\mathbf{w}\| \to \infty$*. Η εξίσωσή του με τη μονάδα είναι απλώς βολική για τους υπολογισμούς μας.*

Χρησιμοποιώντας έναν πολλαπλασιαστή Lagrange λ, το πρόβλημα μεταφράζεται στην εύρεση των στάσιμων σημείων της

$$\Lambda_1(\mathbf{w}, \lambda) = \mathbf{w}^T S\mathbf{w} - \lambda\left(\mathbf{w}^T\mathbf{w} - 1\right) \tag{2.20}$$

Έχουμε λοιπόν

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{w}}\left[\mathbf{w}^T S\mathbf{w} - \lambda\left(\mathbf{w}^T\mathbf{w} - 1\right)\right] = 0 \\ \Rightarrow\ & S\mathbf{w} - \lambda\mathbf{w} = 0 \\ \Rightarrow\ & S\mathbf{w} = \lambda\mathbf{w} \end{aligned} \tag{2.21}$$

και

$$\begin{aligned} &\frac{\partial}{\partial \lambda}\left[\mathbf{w}^T S\mathbf{w} - \lambda\left(\mathbf{w}^T\mathbf{w} - 1\right)\right] = 0 \\ \Rightarrow\ & \mathbf{w}^T\mathbf{w} - 1 = 0 \end{aligned} \tag{2.22}$$

που είναι ο αρχικός περιορισμός μας. Δηλαδή η κατεύθυνση όπου μεγιστοποιείται η διασπορά των δειγμάτων, προσδιορίζεται από ένα ιδιοδιάνυσμα του πίνακα συνδιακύμανσης S. Παρατηρώντας ότι για ένα (κανονικοποιημένο) ιδιοδιάνυσμα $\mathbf{v}_k$ και την αντίστοιχη ιδιοτιμή λ_k ισχύει

$$\mathbf{v}_k^T S\mathbf{v}_k = \mathbf{v}_k^T\lambda_k\mathbf{v}_k = \lambda_k\mathbf{v}_k^T\mathbf{v}_k = \lambda_k \tag{2.23}$$

είναι σαφές ότι η λύση στο πρόβλημα μεγιστοποίησης της εξίσωσης (2.18), δίνεται από το ιδιοδιάνυσμα του S που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή. Κατ' αντίστοιχο τρόπο, όλες οι υπόλοιπες κύριες συνιστώσες $\mathbf{w}_i$ μεγιστοποιούν τη διασπορά υπό τον επιπλέον περιορισμό ότι είναι κάθετες σε όλες τις προηγούμενες, δηλαδή

$$\langle \mathbf{w}_i, \mathbf{w}_j \rangle = \delta_{i,j}, \quad j = 1, \ldots, i-1 \tag{2.24}$$

Συγκεκριμένα, για τη δεύτερη κύρια συνιστώσα έχουμε

$$\mathbf{w}_2 = \underset{\substack{\|\mathbf{w}\|=1 \\ \langle \mathbf{w}_2, \mathbf{w}_1 \rangle = 0}}{\operatorname{argmax}} J(\mathbf{w}) \tag{2.25}$$

και η συνάρτηση Lagrange είναι η

$$\Lambda_2(\mathbf{w}, \lambda, \gamma) = \mathbf{w}^T S \mathbf{w} + \lambda(\mathbf{w}^T \mathbf{w} - 1) + \gamma(\mathbf{w}_1^T \mathbf{w}) \tag{2.26}$$

Οι περιορισμοί εκφράζονται από τις εξισώσεις

$$\begin{aligned} &\frac{\partial}{\partial \lambda} \left[\mathbf{w}^T S \mathbf{w} + \lambda(\mathbf{w}^T \mathbf{w} - 1) + \gamma(\mathbf{w}_1^T \mathbf{w}) \right] = 0 \\ &\Rightarrow \mathbf{w}^T \mathbf{w} - 1 = 0 \end{aligned} \tag{2.27}$$

και

$$\begin{aligned} &\frac{\partial}{\partial \gamma} \left[\mathbf{w}^T S \mathbf{w} + \lambda(\mathbf{w}^T \mathbf{w} - 1) + \gamma(\mathbf{w}_1^T \mathbf{w}) \right] = 0 \\ &\Rightarrow \mathbf{w}_1^T \mathbf{w} = 0 \end{aligned} \tag{2.28}$$

και η ζητούμενη κατεύθυνση από τη

$$\begin{aligned} &\frac{\partial}{\partial \mathbf{w}} \left[\mathbf{w}^T S \mathbf{w} + \lambda(\mathbf{w}^T \mathbf{w} - 1) + \gamma(\mathbf{w}_1^T \mathbf{w}) \right] = 0 \\ &\Rightarrow 2S\mathbf{w} + 2\lambda \mathbf{w} + \gamma \mathbf{w}_1 = 0 \\ &\Rightarrow 2\mathbf{w}_1^T S \mathbf{w} + 2\lambda \mathbf{w}_1^T \mathbf{w} + \gamma \mathbf{w}_1^T \mathbf{w}_1 = 0 \\ &\Rightarrow 2\lambda_1 \underbrace{\mathbf{w}_1^T \mathbf{w}}_{0} + 2\lambda \underbrace{\mathbf{w}_1^T \mathbf{w}}_{0} + \gamma \underbrace{\mathbf{w}_1^T \mathbf{w}_1}_{1} = 0 \\ &\Rightarrow \gamma = 0 \end{aligned} \tag{2.29}$$

δηλαδή επιστρέφουμε στην εξίσωση (2.20) με έναν επιπλέον περιορισμό. Γνωρίζουμε ότι τα ιδιοδιανύσματα ενός πίνακα *είναι* κάθετα μεταξύ τους, αφού αν $\mathbf{v}_i, \mathbf{v}_j$ τα ιδιοδιανύσματα του S που αντιστοιχούν στις ιδιοτιμές λ_i, λ_j

$$\begin{aligned}
\lambda_i\langle \mathbf{v}_i, \mathbf{v}_j\rangle &= \langle \lambda_i \mathbf{v}_i, \mathbf{v}_j\rangle = \langle S\mathbf{v}_i, \mathbf{v}_j\rangle \\
&= \langle \mathbf{v}_i, S^T\mathbf{v}_j\rangle = \langle \mathbf{v}_i, S\mathbf{v}_j\rangle \\
&= \langle \mathbf{v}_i, \lambda_j\mathbf{v}_j\rangle = \lambda_j\langle \mathbf{v}_i, \mathbf{v}_j\rangle \\
&\Rightarrow (\lambda_i - \lambda_j)\langle \mathbf{v}_i, \mathbf{v}_j\rangle = 0 \\
&\Rightarrow \langle \mathbf{v}_i, \mathbf{v}_j\rangle = 0, \quad (\because \lambda_i \neq \lambda_j)
\end{aligned} \tag{2.30}$$

οπότε προφανώς η δεύτερη κύρια συνιστώσα είναι το ιδιοδιάνυσμα που αντιστοιχεί στη δεύτερη μεγαλύτερη ιδιοτιμή. Με τον ίδιο τρόπο βρίσκουμε και τις υπόλοιπες κύριες συνιστώσες και τελικά ο πίνακας για την αλλαγή βάσης που αναφέραμε στην αρχή είναι ο

$$Q = \left[\begin{bmatrix} \\ \mathbf{v}_1 \\ \\ \end{bmatrix} \begin{bmatrix} \\ \mathbf{v}_2 \\ \\ \end{bmatrix} \cdots \begin{bmatrix} \\ \mathbf{v}_d \\ \\ \end{bmatrix} \right] \tag{2.31}$$

με $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d$ τα ιδιοδιανύσματα του S που αντιστοιχούν στις ιδιοτιμές του $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. Υπό την υπόθεση ότι η μεγάλη διασπορά προς μία κατεύθυνση συνεπάγεται πως σε αυτή την κατεύθυνση βρίσκεται το μεγαλύτερο μέρος της πληροφορίας, οι πρώτες συνιστώσες των μετασχηματισμένων παρατηρήσεων $\mathbf{y} = Q\mathbf{x}$ είναι οι πιο σημαντικές. Συγκεκριμένα από τον ορισμό της συνολικής διασποράς

$$\begin{aligned}
\sigma^2_{\text{ολ.}} = tr(S) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_d^2 &= \sum_{i=1}^{d} \mathrm{Var}\left[\mathbf{u_i}\right] \\
= \lambda_1 + \lambda_2 + \cdots + \lambda_d &= \sum_{i=1}^{d} \mathrm{Var}\left[\mathbf{w}_i\right]
\end{aligned} \tag{2.32}$$

όπου $tr(S)$ το ίχνος του πίνακα S, βλέπουμε ότι ο λόγος $\lambda_i/tr(S)$ δίνει το ποσοστό διασποράς των δεδομένων που δικαιολογείται από τη κύρια συνιστώσα $\mathbf{w}_i$.

Μία διαφορετική αλλά ισοδύναμη ερμηνεία της ανάλυσης κύριων συνιστωσών, είναι ότι μέσω της αλλαγής βάσης αφαιρούμε τον πλεονασμό που υπάρχει λόγω συσχέτισης των αρχικών μεταβλητών $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d$. Πράγματι, λαμβάνοντας υπόψη την

παραγοντοποίηση του πίνακα S

$$S = Q\Lambda Q^T \tag{2.33}$$

όπου Λ ο διαγώνιος πίνακας των ιδιοτιμών

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{bmatrix} \tag{2.34}$$

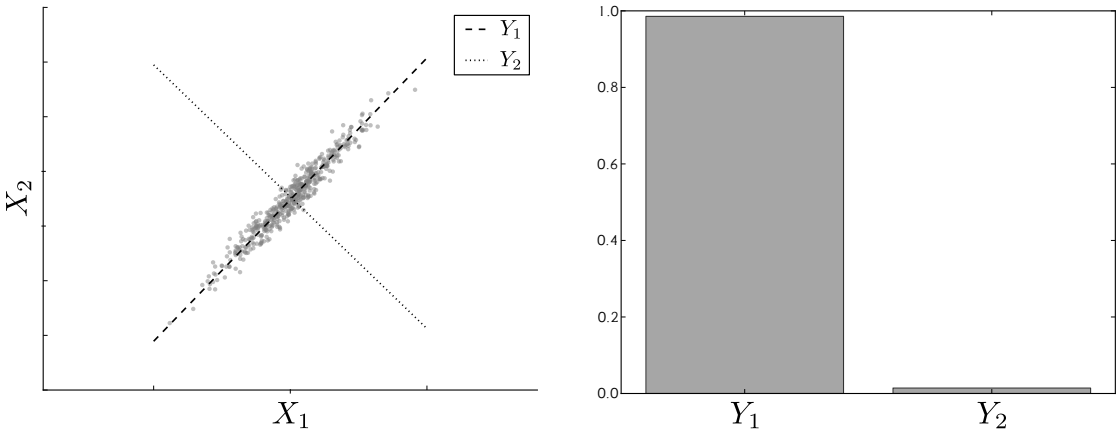
παρατηρούμε ότι μέσω του μετασχηματισμού $\mathbf{y} = Q\mathbf{x}$ έχουμε

$$\begin{aligned} \mathrm{Var}\left[\mathbf{y}\right] = QSQ^T &= Q(Q^T\Lambda Q)Q^T \\ &= (QQ^T)\Lambda(QQ^T) \\ &= (QQ^{-1})\Lambda(QQ^{-1}) \\ &= \Lambda \end{aligned} \tag{2.35}$$

που είναι φυσικά διαγώνιος. Αυτό το αποτέλεσμα υποδεικνύει και τις περιπτώσεις που η ανάλυση κύριων συνιστωσών μπορεί όντως να φανεί χρήσιμη.

Τα παραπάνω μπορούν να γίνουν πιο ξεκάθαρα στην πράξη. Ας πάρουμε για παράδειγμα τα δεδομένα του Σχήματος 2.7. Εμφανώς η συσχέτιση είναι πολύ ισχυρή και η αλλαγή αξόνων στις κύριες συνιστώσες μοιάζει λογική. Όπως αποδείξαμε, οι νέες μεταβλητές είναι ασυσχέτιστες και το ποσοστό της μεταβλητότητας των δεδομένων που δικαιολογεί η καθεμιά από αυτές φαίνεται στο δεξιό μέρος του σχήματος. Θεωρώντας ότι η αρχική μας υπόθεση ισχύει (ότι η χρήσιμη πληροφορία περιέχεται στις μεταβλητές με μεγάλη διασπορά), είναι εύκολο να πάρουμε την απόφαση να διώξουμε τη μεταβλητή Y_2 μειώνοντας έτσι τον όγκο των δεδομένων μας στο μισό.

Το παραπάνω παραδείγμα δίνει μια ιδέα για το τι μπορούμε να πετύχουμε με την Ανάλυση Κύριων Συνιστωσών, αλλά τα αποτελέσματα σε πραγματικά δεδομένα είναι συχνά εντυπωσιακά και δικαιολογούν με καλύτερο τρόπο τους λόγους που αυτή η μέθοδος είναι τόσο διαδεδομένη. Αυτό γίνεται επειδή όπως έχουμε ξαναπεί είναι σύνηθες να χρησιμοποιούνται πολλές μεταβλητές σε κάποιο πρόβλημα και αρκετές φορές η συσχέτιση που παρατηρείται είναι μεγάλη, με αποτέλεσμα να υπάρχει



Σχήμα 2.7: Ανάλυση Κύριων Συντιστωσών σε δεδομένα με μεγάλη συσχέτιση

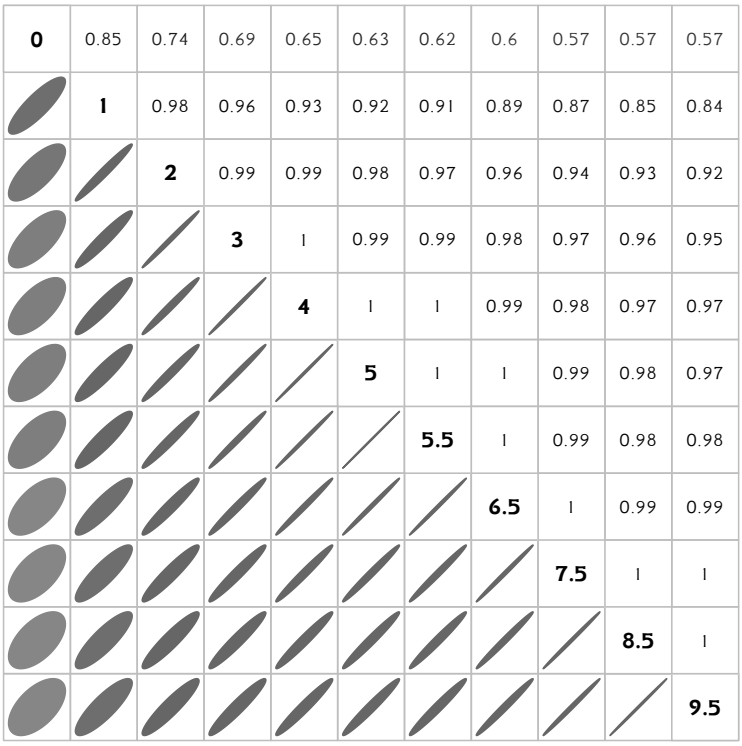
δυνατότητα σημαντικής μείωσής τους. Στο τελευταίο αυτό παράδειγμα, θα χρησιμοποιήσουμε δεδομένα της Αμερικάνικης καμπύλης απόδοσης όπως αυτά συλλέχθηκαν από την τράπεζα Bank of International Settlements [1]. Τα δεδομένα αυτά περιέχουν τις αποδόσεις Αμερικανικών ομολόγων για 1352 διαδοχικές ημέρες και για χρόνους ωρίμανσης από 0 έως 9.5 μήνες (συνολικά 11 διαφορετικοί χρόνοι ωρίμανσης). Με άλλα λόγια, έχουμε 1352 δείγματα διάστασης 11. Όπως φαίνεται στο Σχήμα 2.8 υπάρχει πολύ μεγάλος βαθμός συσχέτισης και συνεπώς η Ανάλυση Κύριων Συνιστωσών αναμένεται να έχει καλά αποτελέσματα όσον αφορά τη μείωση διαστάσεων. Πράγματι μετά την ανάλυση, βλεπουμε ότι το 99.9% της μεταβλητότητας των δεδομένων εξηγείται από τις τρεις πρώτες κύριες συνιστώσες. Σχηματικά αυτό φαίνεται στο Σχήμα 2.9. Όσον αφορά την ερμηνεία αυτού του αποτελέσματος, θα πρέπει να δούμε τα ίδια τα ιδιοδιανύσματα για να ξέρουμε ποιες από τις αρχικές μεταβλητές συμμετέχουν περισσότερο σε αυτές τις κατευθύνσεις, κάτι που δεν μας ενδιαφέρει στο παρόν κεφάλαιο. Το σημαντικό συμπέρασμα είναι ότι μπορούμε μέσα από μία απλή διαδικασία να μειώσουμε δραστικά τον αριθμό μεταβλητών που συμμετέχουν στο πρόβλημά μας, αποβάλλοντας την πλεονάζουσα πληροφορία που περιέχεται συχνά σε πρωτογενή δεδομένα.

[1] Η διάθεση των δεδομένων είναι ελεύθερη στην ηλεκτρονική διεύθυνση https://r-forge.r-project.org/scm/viewvc.php/pkg/Rsafd/man/us.bis.yield.Rd

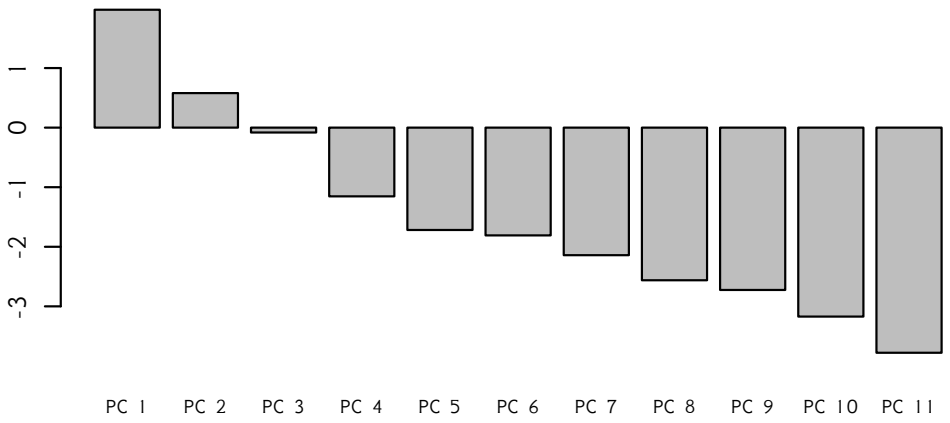
2.2.2 Γραμμική Διαχωριστική Ανάλυση

Στην ανάλυση κύριων συνιστωσών θεωρούμε ότι οι κατευθύνσεις όπου δεν υπάρχει μεγάλη μεταβλητότητα είναι περιττές και έτσι αφαιρούνται. Είναι δηλαδή μια μη-επιβλεπόμενη τεχνική, όπου θεωρείται ότι όλη η πληροφορία που μας ενδιαφέρει περιέχεται στη θέση του κάθε δείγματος στο χώρο $\mathbb{R}^d$. Αντίθετα, στη γραμμική διαχωριστική ανάλυση (**L**inear **D**iscriminant **A**nalysis - LDA) χρησιμοποιούμε δεδομένα που είναι ταξινομημένα σε κλάσεις και επιλέγονται κατευθύνσεις στις οποίες τα δεδομένα διαχωρίζονται, διευκολύνοντας έτσι την ομαδοποίηση στον υπόχωρο που δημιουργείται.

Για αρχή θα δώσουμε κάποιους ορισμούς για την περιγραφή των δεδομένων σε σχέση με τις κλάσεις όπου ανήκουν. Θεωρούμε ότι έχουμε ένα σύνολο N παρατηρήσεων $\mathcal{X} = \left\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right\} \subset \mathbb{R}^d$, ένα σύνολο K κλάσεων $\mathcal{C} = \left\{c_1, \ldots, c_K\right\}$ και ένα σύνολο αντιστοιχήσεων μεταξύ παρατηρήσεων και κλάσεων $\left\{\mathbf{x}_i \mapsto c^i\right\}$. Ορίζουμε τον



Σχήμα 2.8: Συσχετίσεις του συνόλου δεδομένων us.bis.yield



Σχήμα 2.9: Λογάριθμος του ποσοστού μεταβλητότητας που εξηγείται από κάθε κύρια συνιστώσα

πίνακα διασποράς μεταξύ κλάσεων ως

$$S_b = \underset{c \in \mathcal{C}}{\mathcal{E}} \left[\left(\mathbf{m}_c - \mathbf{m} \right) \left(\mathbf{m}_c - \mathbf{m} \right)^T \right] \tag{2.36}$$

όπου

$$\mathbf{m} = \underset{\mathbf{x}_i \in \mathcal{X}}{\mathcal{E}} \left[\mathbf{x}_i \right] \tag{2.37}$$

$$\mathbf{m}_c = \underset{\mathbf{x}_i \mapsto c}{\mathcal{E}} \left[\mathbf{x}_i \right] \tag{2.38}$$

Επίσης, ορίζουμε τον *πίνακα διασποράς εντός κλάσεων* ως

$$S_w = \underset{c \in \mathcal{C}}{\mathcal{E}} \left[\underset{\mathbf{x}_i \mapsto c}{\mathcal{E}} \left[\left(\mathbf{x}_i - \mathbf{m}_c \right) \left(\mathbf{x}_i - \mathbf{m}_c \right)^T \right] \right] \tag{2.39}$$

και το *συνολικό (ή μικτό) πίνακα διασποράς* των δεδομένων ως

$$S_m = \underset{\mathbf{x}_i \in \mathcal{X}}{\mathcal{E}} \left[\left(\mathbf{x}_i - \mathbf{m} \right) \left(\mathbf{x}_i - \mathbf{m} \right)^T \right] \tag{2.40}$$

Θα ξεκινήσουμε την περιγραφή της μεθόδου, με την πιο απλή περίπτωση όπου υπάρχουν δύο κλάσεις c_1 και c_2 και ψάχνουμε τη "βέλτιστη" κατεύθυνση $\mathbf{w}$ για

το διαχωρισμό τους. Είναι εύκολο να δούμε ότι αν $\mathbf{m}_j = \mathcal{E}_{\mathbf{x}_i \mapsto c_j}\left[\mathbf{x}_i\right]$ και $S_j = \mathcal{E}_{\mathbf{x}_i \mapsto c_j}\left[\left(\mathbf{x}_i - \mathbf{m}_j\right)\left(\mathbf{x}_i - \mathbf{m}_j\right)^T\right]$, τότε για τις προβολές των στοιχείων του $\mathcal{X}$ στη διεύθυνση ενός διανύσματος $\mathbf{w}$, $\mathbf{y}_i = \mathbf{w}^T\mathbf{x}$ έχουμε

$$\begin{aligned} \tilde{m}_j &= \underset{y_i \mapsto c_j}{\mathcal{E}}\left[y_i\right] = \underset{\mathbf{x}_i \mapsto c_j}{\mathcal{E}}\left[\mathbf{w}^T\mathbf{x}_i\right] \\ &= \mathbf{w}^T\mathbf{m}_j, \quad j = 1, 2 \end{aligned} \tag{2.41α'}$$

$$\begin{aligned} \tilde{s}_j^2 &= \underset{y_i \mapsto c_j}{\mathcal{E}}\left[(y_i - \tilde{m}_j)^2\right] \\ &= \underset{\mathbf{x}_i \mapsto c_j}{\mathcal{E}}\left[(\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{m}_j)^2\right] \\ &= \mathbf{w}^T S_j \mathbf{w}, \quad j = 1, 2 \end{aligned} \tag{2.41β'}$$

Η κατεύθυνση που αναζητούμε δίνεται μέσω του *κριτηρίου Fisher*, δηλαδή μεγιστοποιώντας τη συνάρτηση διαχωρισμού

$$J(\mathbf{w}) = \frac{\left(\tilde{m}_1 - \tilde{m}_2\right)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \tag{2.42}$$

Η ερμηνεία αυτής της συνάρτησης είναι προφανής: στον αριθμητή δηλώνεται ότι θέλουμε να υπάρχει μεγάλη απόσταση μεταξύ των κλάσεων και στον παρονομαστή ότι θέλουμε η κάθε κλάση να είναι "συμπαγής" (να έχει μικρή διασπορά). Χρησιμοποιώντας τις εξισώσεις (2.41α') και (2.41β') έχουμε

$$\begin{aligned} J(\mathbf{w}) &= \frac{\left(\tilde{m}_1 - \tilde{m}_2\right)^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \\ &= \frac{\left(\mathbf{w}^T\mathbf{m}_1 - \mathbf{w}^T\mathbf{m}_2\right)^2}{\mathbf{w}^T S_1 \mathbf{w} + \mathbf{w}^T S_2 \mathbf{w}} \\ &= \frac{\left(\mathbf{w}^T\left(\mathbf{m}_1 - \mathbf{m}_2\right)\right)^2}{\mathbf{w}^T\left(S_1 + S_2\right)\mathbf{w}} \\ &\propto \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}} \end{aligned} \tag{2.43}$$

Για τη μεγιστοποίησή της (2.43) θέτουμε

$$\begin{aligned}
\frac{d}{d\mathbf{w}}J(\mathbf{w}) &\propto \frac{\left(\frac{d}{d\mathbf{w}}\mathbf{w}^T S_b \mathbf{w}\right)\mathbf{w}^T S_w \mathbf{w} - \left(\frac{d}{d\mathbf{w}}\mathbf{w}^T S_w \mathbf{w}\right)\mathbf{w}^T S_b \mathbf{w}}{\left(\mathbf{w}^T S_w \mathbf{w}\right)^2} \\
&= \frac{(2S_b\mathbf{w})\mathbf{w}^T S_w \mathbf{w} - (2S_w\mathbf{w})\mathbf{w}^T S_b \mathbf{w}}{\left(\mathbf{w}^T S_w \mathbf{w}\right)^2} = 0 \\
&\Rightarrow \mathbf{w}^T S_w \mathbf{w}\left(S_b\mathbf{w}\right) - \mathbf{w}^T S_b \mathbf{w}\left(S_w\mathbf{w}\right) = 0 \\
&\Rightarrow S_b\mathbf{w} - \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}\left(S_w\mathbf{w}\right) = 0 \\
&\Rightarrow S_b\mathbf{w} = \lambda S_w\mathbf{w}, \quad \text{με } \lambda = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}
\end{aligned} \tag{2.44}$$

δηλαδή η βέλτιστη κατεύθυνση $\mathbf{w}$ προσδιορίζεται από την επίλυση ενός γενικευμένου προβλήματος ιδιοτιμών ή (στην περίπτωση που υπάρχει ο S_w^{-1}) του απλού προβλήματος ιδιοτιμών

$$S_w^{-1} S_b \mathbf{w} = \lambda \mathbf{w} \tag{2.45}$$

και την επιλογή του ιδιοδιανύσματος που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή.

Η γενίκευση της τεχνικής σε περισσότερες διαστάσεις γίνεται με παρόμοια κριτήρια. Πλέον όμως δεν αναζητούμε μία κατεύθυνση $\mathbf{w}$ αλλά έναν υπόχωρο διάστασης $d' < d$. Αφού αποφασιστεί η διάσταση d', το πρόβλημα ανάγεται στον προσδιορισμό ενός $d \times d'$ πίνακα $\hat{A}$ που ικανοποιεί ένα κριτήριο της μορφής

$$\hat{A} = \underset{A \in \mathbb{R}^{d \times d'}}{\operatorname{argmax}}\, J(A, S_1, S_2) \tag{2.46}$$

όπου (S_1, S_2) είναι ένα ζεύγος πινάκων που εξυπηρετούν τους σκοπούς της ανάλυσης, όπως π.χ (S_b, S_w), (S_m, S_w), (S_b, S_m).

Το πιο σύνηθες κριτήριο είναι το *γενικευμένο κριτήριο Fisher*

$$J_1(\mathrm{A}) = tr\left(\left(A^T S_2 A\right)^{-1}\left(A^T S_1 A\right)\right) \tag{2.47}$$

το οποίο ανάγεται όπως και στην απλή περίπτωση που παρουσιάσαμε πριν, στην επίλυση του γενικευμένου προβλήματος ιδιοτιμών

$$S_1\mathbf{w} = \lambda S_2 \mathbf{w} \tag{2.48}$$

και ο πίνακας $\hat{A} = \operatorname{argmax} J(A)$ έχει τη μορφή

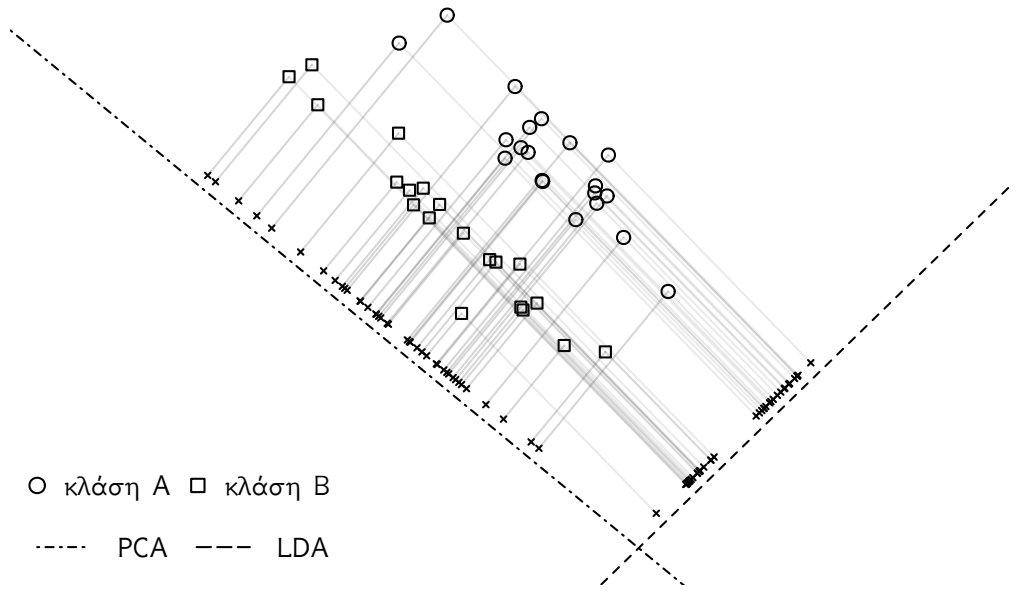
$$\hat{A} = \left[\begin{bmatrix} \\ \mathbf{w}_1 \\ \\ \end{bmatrix} \begin{bmatrix} \\ \mathbf{w}_2 \\ \\ \end{bmatrix} \cdots \begin{bmatrix} \\ \mathbf{w}_{d'} \\ \\ \end{bmatrix} \right] \tag{2.49}$$

όπου $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{d'}$ τα ιδιοδιανύσματα που αντιστοιχούν στις d' μεγαλύτερες ιδιοτιμές.

Ένα ακόμα συνηθισμένο κριτήριο είναι το *κριτήριο λόγου ιχνών*

$$J_2(\mathrm{A}) = \frac{tr\left(A^T S_1 A\right)}{tr\left(A^T S_2 A\right)} \tag{2.50}$$

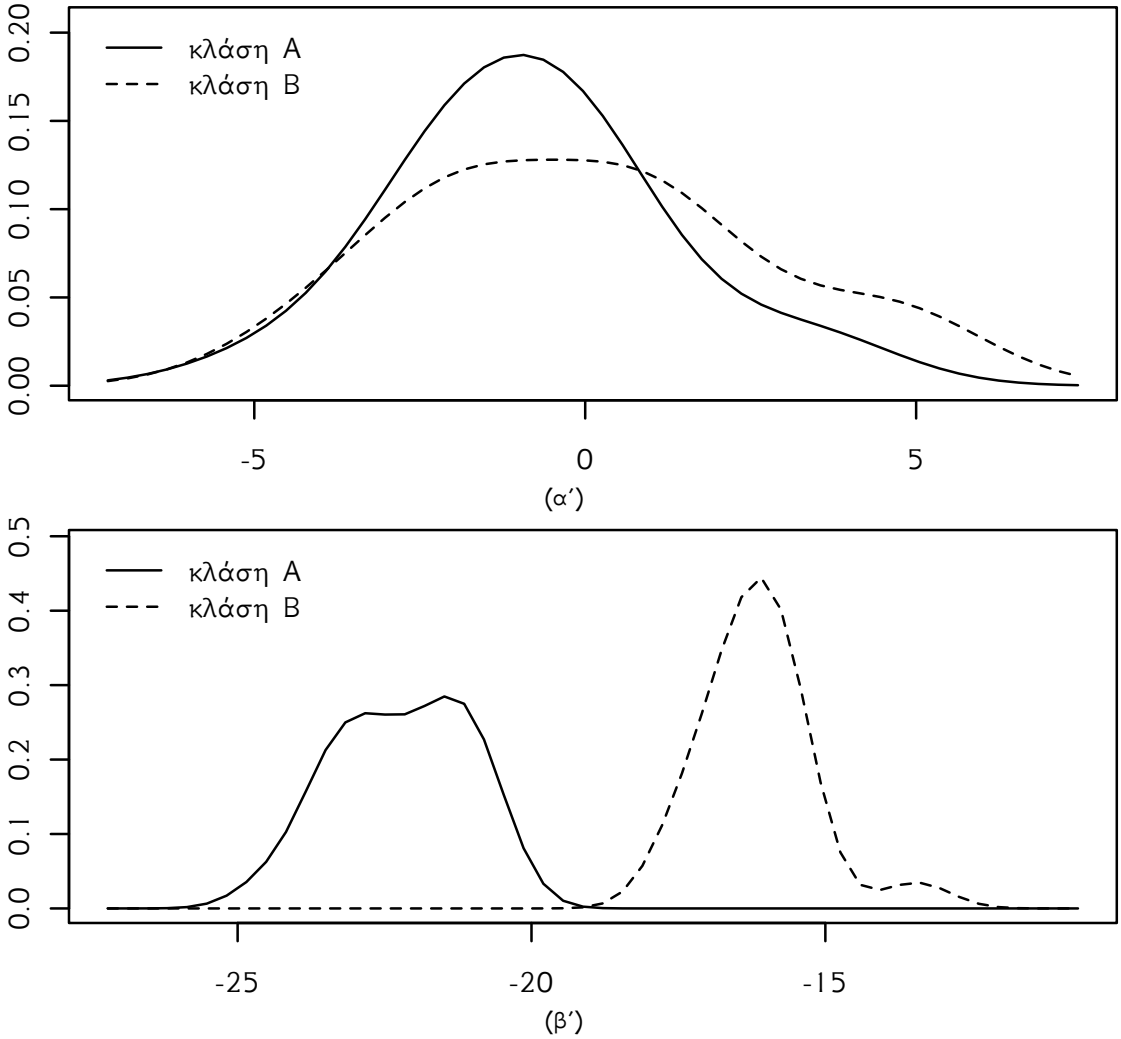
το οποίο μελετήθηκε πρώτη φορά από τους Foley και Sammon στο [Foley and Sammon, 1975] αλλά δεν έχει λύση σε κλειστή μορφή. Παρ' όλα αυτά, υπάρχουν αποδοτικές μέθοδοι επίλυσής του (π.χ. [Zhang et al., 2013]) που επιτρέπουν τη χρήση του σε απαιτητικές εφαρμογές.



Σχήμα 2.10

Είναι σαφές ότι οι δύο προσεγγίσεις στη μείωση διαστάσεων που έχουμε περιγράφει είναι τελείως διαφορετικές και συχνά αυτό αποτυπώνεται στα αποτελέσματά τους.

Ένα χαρακτηριστικό παράδειγμα δίνεται στο Σχήμα 2.10 όπου οι ευθείες προβολής που επιλέγονται από κάθε μέθοδο είναι σχεδόν κάθετες. Βλέπουμε ότι αν δεν ληφθούν υπόψη οι κλάσεις, ο διαχωρισμός των δεδομένων είναι αδύνατος (Σχήμα 2.11α′), ενώ αντίθετα μέσω γραμμικής διαχωριστικής ανάλυσης είναι πολύ εύκολος (Σχήμα 2.11β′).



Σχήμα 2.11

Αυτό δεν σημαίνει ότι οι δύο τεχνικές είναι "ασύμβατες". Μάλιστα σε πολλές εφαρμογές όπου υπάρχει πολύ μεγάλος αριθμός διαστάσεων και σχετικά λίγα δείγματα, η "κατάρα των διαστάσεων" καθιστά αδύνατη τη χρήση γραμμικής διαχωριστικής ανάλυσης εξαιτίας της μη-αντιστρεψιμότητας του πίνακα S_w. Ας παρατηρήσουμε ότι

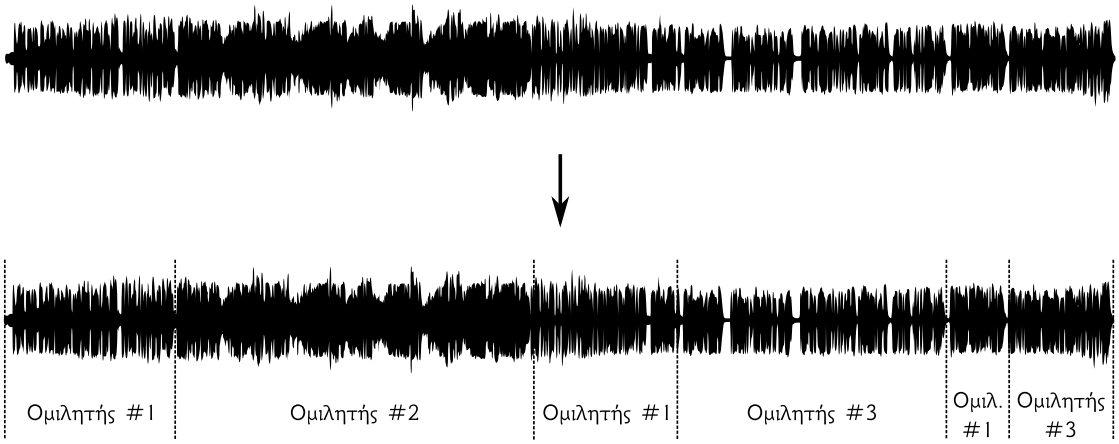
για το διαχωρισμό K ομάδων χρειαζόμαστε $K-1$ διαστάσεις που είναι και ο βαθμός του πίνακα S_b. Για την αντιμετώπιση αυτού του προβλήματος συχνά γίνεται ένας συνδυασμός των δύο τεχνικών. Για παράδειγμα, σε πολλές εργασίες (π.χ [Belhumeur et al., 1997], [Dagher, 2010]) έχουν προταθεί μέθοδοι όπου προηγείται μια διαδικασία ανάλυσης κύριων συνιστωσών και ακολουθεί η γραμμική διαχωριστική ανάλυση στο χώρο της ήδη μειωμένης διάστασης. Μία άλλη προσέγγιση είναι ή προβολή αρχικά στο μηδενικό χώρο του S_w όπου έχει αναφερθεί [Chen, 2000] (αν και δεν είναι καθολικά αποδεκτό) ότι περιέχει την περισσότερη πληροφορία για ταξινόμηση και στη συνέχεια η μείωση σε $K-1$ διαστάσεις μέσω ανάλυσης κύριων συνιστωσών.

ΚΕΦΑΛΑΙΟ 3

Διαχωρισμός Ομιλητών

Ο διαχωρισμός ομιλητών είναι μία από τις βασικές εργασίες στο ευρύ πεδίο της επεξεργασίας ομιλίας. Το αντικείμενό του είναι η απάντηση στο ερώτημα *"ποιός μίλησε και πότε;"*. Με άλλα λόγια, παίρνοντας ένα σήμα ομιλίας, προσπαθούμε να παράξουμε μία κατάτμησή του σε ομοιογενή ως προς τον ομιλητή τμήματα, αλλά και να ταυτοποιήσουμε τα τμήματα που προέρχονται από τον ίδιο ομιλητή (Σχ. 3.1). Σημειώνουμε ότι ο *διαχωρισμός* ομιλητών είναι διαφορετική εργασία από την *αναγνώριση* ομιλητή όπου προσπαθούμε να βρούμε τα τμήματα που μίλησε ένα συγκεκριμένο (και γνωστό εκ των προτέρων) άτομο. Στο διαχωρισμό, οι ομιλητές παραμένουν άγνωστοι (π.χ. "Ομιλητής # 1"). Τα οφέλη από μία τέτοια διαδικασία είναι πολλαπλά. Κατ' αρχάς, ο διαχωρισμός από μόνος του βρίσκει εφαρμογές στην αποθήκευση, την αναζήτηση και την ανάσυρση τμημάτων ομιλίας, κάτι που είναι πολύ χρονοβόρο όταν γίνεται "στα τυφλά", ειδικά αν αναλογιστούμε τον όγκο ηχογραφήσεων ομιλίας που είναι διαθέσιμος σήμερα, ή ακόμα και στη μελέτη της δομής ενός διαλόγου. Αν συνεχίσουμε λίγο ακόμα αυτή τη σκέψη, μπορούμε να φανταστούμε και πιο περίπλοκες εφαρμογές, όπως για παράδειγμα τον αυτόματο υποτιτλισμό εκπομπών, την απομαγνητοφώνηση συνεντεύξεων, την καταγραφή πρακτικών κτλ. απο όπου ο διαχωρισμός ομιλητών είναι μέρος ενός μεγαλύτερου συστήματος.

Για την ανάπτυξη μεθόδων διαχωρισμού ομιλητών είναι απαραίτητες κάποιες υποθέσεις σχετικές με το σήμα ομιλίας:



Σχήμα 3.1: Διαχωρισμός Ομιλητών σε ένα απόσπασμα δελτίου ειδήσεων

1. **Το σήμα περιέχει *μόνο* ομιλία,**
   Αυτή η υπόθεση δεν είναι πάντα ρεαλιστική, αφού συχνά το περιεχόμενο ενός ηχογραφημένου σήματος περιέχει μουσική, παρασκηνιακό θόρυβο κτλ., αλλά είναι συχνό η εργασία διαχωρισμού ομιλητών να γίνεται στα πλαίσια συστημάτων όπου γίνεται κάποια προεπεξεργασία του σήματος για την αντιμετώπιση αυτών των προβλημάτων. Οι μέθοδοι που εφαρμόζονται ανήκουν στην κατηγορία της *Ανίχνευσης Φωνητικής Δραστηριότητας*, που από μόνη της αποτελεί ένα σημαντικό πεδίο έρευνας με ευρύτατο πεδίο εφαρμογών όπως στην κωδικοποίηση και μετάδοση ομιλίας (π.χ. σε εφαρμογές VOIP). Στην παρούσα εργασία δεν ασχολούμαστε με αυτό, αλλά σε όσα ακολουθούν θα θεωρούμε ότι το εκάστοτε σήμα ικανοποιεί την υπόθεσή μας.

2. **Σε κάθε στιγμή μιλάει ακριβώς ένας ομιλητής,**
   Με άλλα λόγια, υποθέτουμε ότι δεν υπάρχουν τμήματα του σήματος όπου δύο ή περισσότεροι ομιλητές μιλάνε ταυτόχρονα. Στην πραγματικότητα βέβαια, αυτό συμβαίνει, αλλά η διάρκεια αυτών των τμημάτων είναι συνήθως αμελητέα σε σχέση με τη διάρκεια του σήματος, συνεπώς μπορούμε να υποθέσουμε ότι η υπόθεσή μας ισχύει.

Κατά καιρούς έχουν προταθεί διάφορα χαρακτηριστικά του σήματος ομιλίας ως κατάλληλα για διαχωρισμό ομιλητών, όπως οι συντελεστές LSP [Adami et al., 2002],

οι σαφματικοί συντελεστές γραμμικής πρόβλεψης LPCC [Ajmera and Wooters, 2003] και οι σαφματικοί συντελεστές mel κλίμακας MFCC [Zhou and Hansen, 2005] με τους τελευταίους να αποτελούν τη πιο συνήθη επιλογή. Γενικά, τα αποτελέσματα που προκύπτουν με την επιλογή αυτών των χαρακτηριστικών είναι παρόμοια και ικανοποιητικά. Ο τομέας όπου δίνεται έμφαση είναι *ο τρόπος* με τον οποίο μπορούν να χρησιμοποιηθούν για το διαχωρισμό ομιλητών και με αυτό θα ασχοληθούμε στη συνέχεια.

3.1 Τυπικές Μέθοδοι Διαχωρισμού Ομιλητών

Η παραδοσιακή προσέγγιση στο διαχωρισμό ομιλητών περιλαμβάνει δύο στάδια: σε μία πρώτη φάση εντοπίζονται τα σημεία αλλαγής ομιλητή και στη συνέχεια τμήματα που προκύπτουν συγκρίνονται για να διαπιστωθεί αν ανήκουν στον ίδιο ή διαφορετικό ομιλητή.

Το κοινό σημείο των παραδοσιακών μεθόδων είναι η προσαρμογή ενός στατιστικού μοντέλου σε ένα τμήμα της χρονοσειράς χαρακτηριστικών ενός τμήματος ομιλίας. Στη βιβλιογραφία έχουν προταθεί πολλλές μέθοδοι επιλογής μοντέλου παραμετρικές και μη παραμετρικές, με την πιο δημοφιλή να είναι το Κριτήριο Μπεϋζιανής Πληροφορίας (**B**ayesian **I**nformation **C**riterion - BIC) [Schwarz, 1978] που αποτελεί ένα κριτήριο πιθανοφάνειας με ένα συντελεστή ποινής στον αριθμό παραμέτρων του μοντέλου.

Συγκεκριμένα, ας συμβολίσουμε με $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ το σύνολο παρατηρήσεων και $\mathcal{M} = \{M_1, M_2, \ldots, M_K\}$ το σύνολο των υποψήφιων παραμετρικών μοντέλων. Υποθέτουμε ότι η συνάρτηση πιθανοφάνειας του κάθε μοντέλου M έχει μεγιστοποιηθεί ξεχωριστά για τις παρατηρήσεις στην ποσότητα $L(\mathcal{X}, M)$. Το Κριτήριο Μπεϋζιανής Πληροφορίας ορίζεται ως

$$BIC(M) = \log L(\mathcal{X}, M) - \lambda \frac{1}{2} \#(M) \times \log(N) \tag{3.1}$$

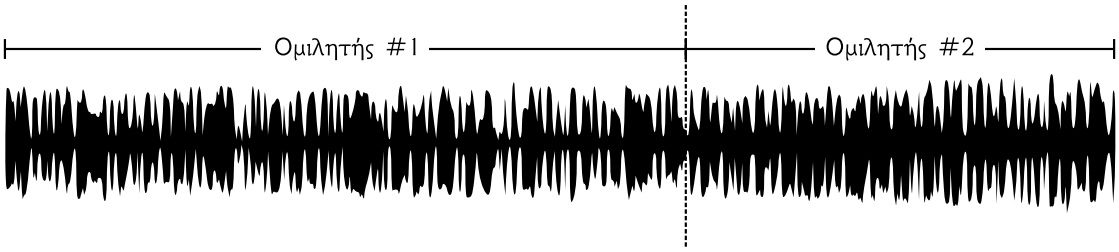
όπου $\#(M)$ είναι ο αριθμός παραμέτρων του μοντέλου M, ενώ η τιμή του συντελεστή ποινής λ στον ορισμό είναι $\lambda = 1$ (αν και μπορεί να αλλάξει).

Για να δούμε πως χρησιμοποιείται το κριτήριο BIC στο διαχωρισμό ομιλητών, παρουσιάζουμε τη *μέθοδο ανίχνευσης ενός σημείου αλλαγής* [Chen and Gopalakrishnan,

1998] που μπορεί να θεωρηθεί (ως προς τη φιλοσοφία της τουλάχιστον) σαν κοινός παρονομαστής μεγάλου μέρους των υπόλοιπων παραδοσιακών μεθόδων.

3.1.1 Ανίχνευση ενός σημείου αλλαγής

Υποθέτουμε ότι το σήμα που επεξεργαζόμαστε περιέχει *το πολύ* ένα σημείο αλλαγής ομιλητή όπως στο Σχήμα 3.2.



Σχήμα 3.2: Απόσπασμα δελτίου ειδήσεων με ένα σημείο αλλαγής ομιλητή

Επίσης, θεωρούμε ότι κάθε παρατήρηση $\mathbf{x}_i \in \mathcal{X}$ ακολουθεί πολυμεταβλητή κανονική κατανομή

$$\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{3.2}$$

Η υπόθεση ύπαρξης ενός σημείου αλλαγής της Γκαουσσιανής χρονοσειράς $\{\mathbf{x}_i\}$ στο σημείο $l \in (1, N)$ εκφράζεται από τον έλεγχο υποθέσεων

$$H_0 : \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\mathrm{N} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3.3α'}$$

$$\begin{aligned} H_1 : &\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \\ &\mathbf{x}_{l+1}, \mathbf{x}_{l+2}, \ldots, \mathbf{x}_\mathrm{N} \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \end{aligned} \tag{3.3β'}$$

Στη μηδενική υπόθεση θεωρούμε ότι όλες οι παρατηρήσεις προέρχονται από την ίδια κατανομή, ενώ στην εναλλακτική θεωρούμε ότι οι l πρώτες προέρχονται από μία κατανομή και οι υπόλοιπες $N - l$ από μία άλλη. Στην ουσία έχουμε ένα πρόβλημα

επιλογής μοντέλου με

$$\begin{aligned} BIC(\mathcal{X}, H_0) &= \log p(\mathcal{X};\ \mathbf{m}, S) - \lambda \frac{1}{2} \#(H_0) \log N \\ &= \log \left(\frac{1}{(2\pi)^{dN/2} |S|^{N/2}} e^{-dN/2} \right) - \lambda \frac{1}{2} \left(d + \frac{1}{2} d(d+1) \right) \log N \\ &= -\frac{dN}{2}(1 + \log(2\pi)) - \frac{N}{2} \log |S| - \lambda \frac{1}{2} \left(d + \frac{1}{2} d(d+1) \right) \log N \end{aligned} \tag{3.4α'}$$

και θέτοντας $\mathcal{X}_1 = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i\}$ και $\mathcal{X}_2 = \{\mathbf{x}_{i+1}, \mathbf{x}_{i+2}, \ldots, \mathbf{x}_N\}$

$$\begin{aligned} BIC(\mathcal{X}, H_1) &= \log p\left(\mathcal{X}_1;\ \mathbf{m}_1,\ S_1\right) + \log p\left(\mathcal{X}_2;\ \mathbf{m}_2,\ S_2\right) - \lambda \frac{1}{2} \#(H_1) \log N \\ &= \log \left(\frac{1}{(2\pi)^{di/2} \left|S_1\right|^{i/2}} e^{-di/2} \right) + \log \left(\frac{1}{(2\pi)^{d(N-i)/2} \left|S_2\right|^{(N-i)/2}} e^{-d(N-i)/2} \right) \\ &\quad - \lambda \frac{1}{2} (2d + d(d+1)) \log N \\ &= -\frac{dN}{2}(1 + \log(2\pi)) - \frac{i}{2} \log \left|S_1\right| - \frac{N-i}{2} \log \left|S_2\right| \\ &\quad - \lambda \frac{1}{2} (2d + d(d+1)) \log N \end{aligned} \tag{3.4β'}$$

όπου S, S_1, S_2 οι δειγματικοί πίνακες συνδιακύμανσης των συνόλων $\mathcal{X}$, $\mathcal{X}_1$ και $\mathcal{X}_2$ αντίστοιχα.

Ορίζοντας

$$\begin{aligned} \Delta BIC(i) &= BIC(\mathcal{X}, H_0) - BIC(\mathcal{X}, H_1) \\ &= R(i) - \lambda P \end{aligned} \tag{3.5}$$

με

$$R(i) = N \log |S| - i \log \left|S_1\right| - (N-i) \log \left|S_2\right| \tag{3.6}$$

και

$$P = \frac{1}{2} \left(d + \frac{1}{2} d(d+1) \right) \tag{3.7}$$

η απόφαση για την επιλογή μοντέλου γίνεται μέσω της σχέσης

$$\Delta BIC \underset{H_1}{\overset{H_0}{\gtrless}} 0 \tag{3.8}$$

Μετά από υπολογισμό του Κριτηρίου Μπεϋζιανής Πληροφορίας για $i = 2, 3, \ldots, N-1$, το υποψήφιο σημείο αλλαγής ομιλητή δίνεται απλά από τη σχέση

$$\hat{t} = \underset{i}{\operatorname{argmax}}\, \Delta BIC(i) \tag{3.9}$$

ενώ για να γίνει δεκτό σαν σημείο αλλαγής θα πρέπει φυσικά να ισχύει

$$\max_{i} \Delta BIC(i) > 0 \tag{3.10}$$

Σε αντίθετη περίπτωση θεωρούμε ότι υπάρχει μόνο ένας ομιλητής στο σήμα που εξετάζουμε.

Στην πράξη φυσικά δεν εξετάζεται κάθε σημείο i, αλλά ο δείκτης αυξάνεται με ένα σταθερό βήμα Δ που αντιστοιχεί σε ένα μικρό χρονικό διάστημα δ (π.χ $200-500$ ms) για λόγους εξοικονόμησης υπολογιστικών πόρων, αφού ο υπολογισμός του κριτηρίου απαιτεί τον υπολογισμό δύο πλήρων πινάκων συνδιακύμανσης μεγάλου συνήθως βαθμού και της ορίζουσάς τους. Αυτό γίνεται χωρίς να υπάρχει κάποια σημαντική απώλεια αν σκεφτούμε ότι το χρονικό διάστημα μεταξύ δύο διαδοχικών παρατηρήσεων είναι αμελητέο (τυπικά περίπου 10-20 ms) ενώ μία ακρίβεια της τάξης του μισού δευτερολέπτου είναι επαρκής για τις συνήθεις εφαρμογές. Σαν παράδειγμα, στο Σχήμα 3.3 βλέπουμε τις τιμές της διαφοράς ΔBIC για το σήμα του Σχήματος 3.2. με το σημείο αλλαγής να είναι εμφανές.

Αλγόριθμος 4 Ανίχνευση ενός σημείου αλλαγής $OCD\left(\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}, \delta\right)$

Είσοδος: Η ακολουθία παρατηρήσεων $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$
Το βήμα μετατόπισης ορίου δ

Έξοδος: Το σύνολο σημείων αλλαγής CP

1: Αρχικοποίηση $i = 0$
2: **για όσο** $i < b - \delta$
3: Μετατόπιση ορίου: $i \leftarrow i + \delta$
4: Καθορισμός συνόλων $X_1 = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i\}$, $\mathcal{X}_2 = \{\mathbf{x}_{i+1}, \mathbf{x}_{i+2}, \ldots, \mathbf{x}_N\}$

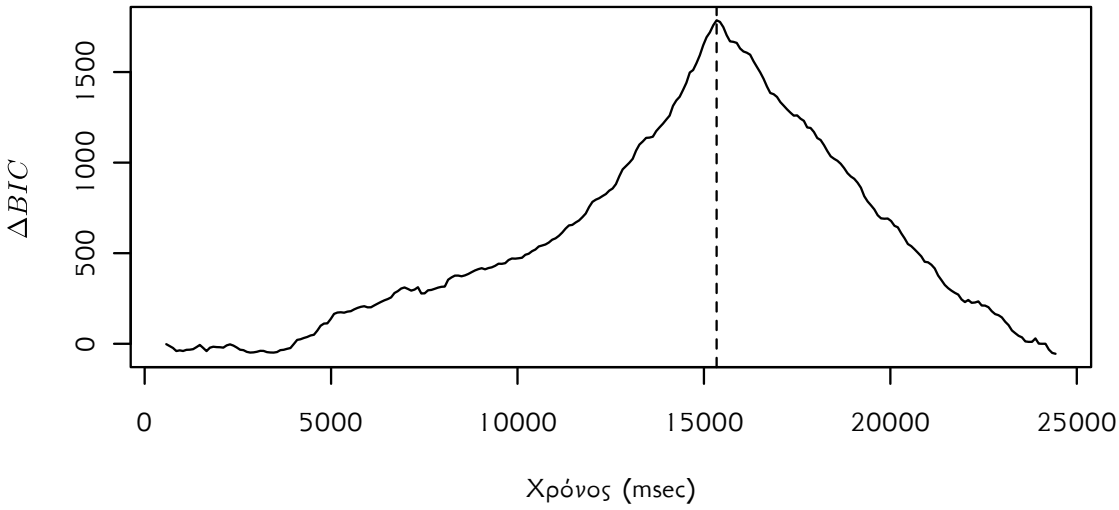
5: Υπολογισμός $\Delta BIC(i)$ από την εξίσωση (3.5)

6: **εάν** $max_i \Delta BIC(i) > 0$ **τότε**

7: $CP = \left\{\text{argmax}_i \, \Delta BIC(i)\right\}$

8: **αλλιώς**

9: $CP = \emptyset$



Σχήμα 3.3: Καμπύλη ΔBIC σήματος με ένα σημείο αλλαγής

3.1.2 Ανίχνευση πολλών σημείων αλλαγής

Η υπόθεση ύπαρξης ενός σημείου αλλαγής είναι φυσικά μη ρεαλιστική για τα σήματα ομιλίας που επεξεργαζόμαστε στην πράξη. Ο Αλγόριθμος 5 αποτελεί μία άμεση προέκταση της προηγούμενης μεθόδου για ανίχνευση πολλών σημείων αλλαγής. Συνοπτικά, σε αυτή τη μέθοδο χρησιμοποιείται ένα παράθυρο όπου πιθανόν να υπάρχει ένα σημείο αλλαγής, το οποίο και αναζητείται με χρήση ενός σχετικά μεγάλου βήματος. Αν δεν βρεθεί, το παράθυρο μεγαλώνει διαδοχικά, όμως μέχρι ενός σημείου. Αν και πάλι δεν έχει βρεθεί σημείο αλλαγής, το όλο παράθυρο μετατοπίζεται μέχρις ότου να βρεθεί κάποιο ή να φτάσουμε στο τέλος του σήματος. Σε όλη τη διάρκεια

Αλγόριθμος 5 Ανίχνευση πολλών σημείων αλλαγής [Tritschler and Gopinath, 1999]

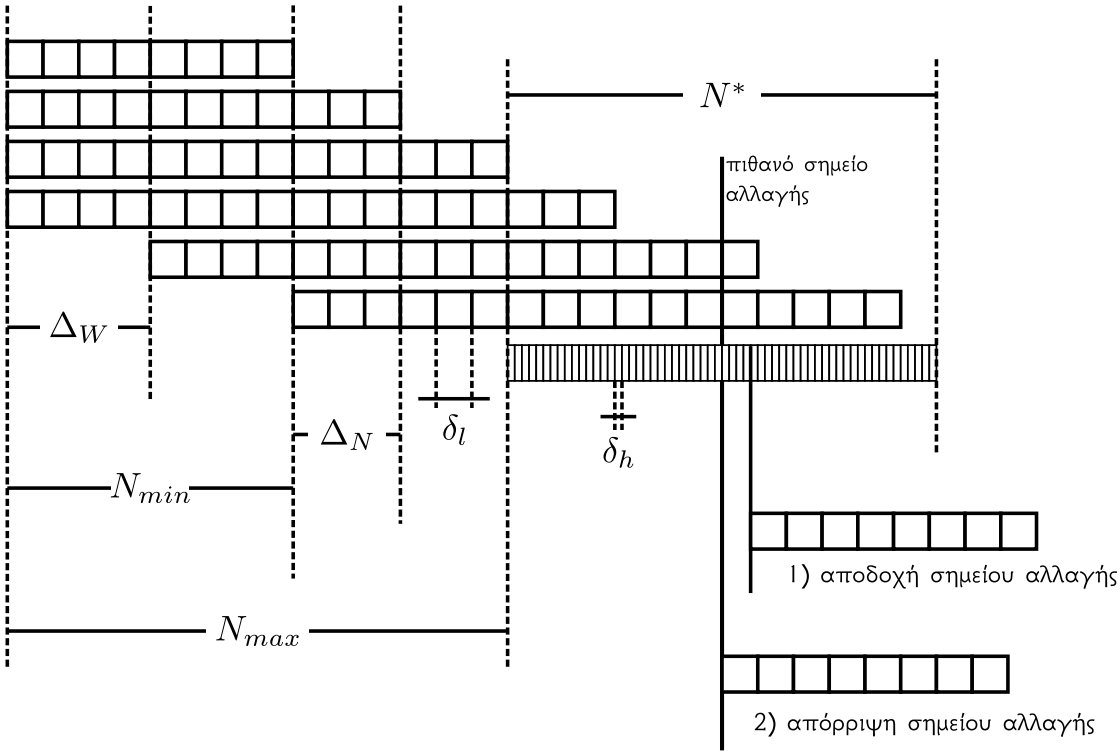
Είσοδος: Η ακολουθία παρατηρήσεων $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$
Τα βήματα μετατόπισης ορίου δ_l, δ_h; $(\delta_h < \delta_l)$
Τα βήματα αύξησης και μετατόπισης παραθύρου Δ_N, Δ_W αντίστοιχα
Τα μεγέθη ελάχιστου, μέγιστου και βοηθητικού παραθύρου N_{min}, N_{max}, N^* αντίστοιχα $(N_{min} \leq N^* \leq N_{max})$

Έξοδος: Το σύνολο σημείων αλλαγής CP

1: Αρχικοποίηση $a = 1$, $b = N_{min}$, $i = 0$, $CP = \emptyset$
2: **για όσο** $OCD\left(\{\mathbf{x}_a, \mathbf{x}_{a+1}, \ldots, \mathbf{x}_b\}, \delta_l\right) = \emptyset$ και $b - a < N_{max} - \Delta_N$
3: $\quad b \leftarrow b + \Delta_N$
4: **εάν** $CP^* = OCD\left(\{\mathbf{x}_a, \ldots, \mathbf{x}_b\}, \delta_l\right) \neq \emptyset$ **τότε**
5: $\quad a^* = t^* - N^*/2$, $b^* = t^* + N^*/2$; $t^* \in CP^*$
6: $\quad CP^{**} = OCD\left(\{\mathbf{x}_{a^*}, \mathbf{x}_{a^*+1}, \ldots, \mathbf{x}_{b^*}\}, \delta_h\right)$
7: $\quad CP \leftarrow CP \cup CP^{**}$
8: $\quad$ **εάν** $CP^{**} = \emptyset$ **τότε**
9: $\quad\quad a \leftarrow t^*$, $b \leftarrow t^* + N_{min}$
10: $\quad\quad$ **πήγαινε στο** 2
11: $\quad$ **αλλιώς**
12: $\quad\quad a \leftarrow t^{**}$, $b \leftarrow a + N_{min}$; $t^{**} \in CP^{**}$
13: $\quad\quad$ **πήγαινε στο** 2
14: **εάν** $b - a \geq N_{max}$ **τότε**
15: $\quad a \leftarrow a + \Delta_W$, $b \leftarrow b + \Delta_W$
16: $\quad$ **πήγαινε στο** 2

αναζήτησης σημείων αλλαγής, το μέγεθος του παραθύρου είναι περιορισμένο και ως προς το ελάχιστο (ώστε να μπορεί να γίνει εκτίμηση των στατιστικών μεγεθών) και ως προς το μέγιστο (ώστε να υπάρχει το πολύ ένα σημείο αλλαγής). Ειδικά το μέγιστο μέγεθος παραθύρου αποτελεί ένα είδος "προστασίας" από τις αστοχίες που μπορεί να υπάρχουν κατά την εκτέλεση της μεθόδου, αφού αν κάποιο σημείο αλλαγής δεν εντοπιστεί, η απόδοσή της μπορεί να μειωθεί πολύ γρήγορα. Τέλος, εάν βρεθεί ένα σημείο αλλαγής, η ύπαρξή του επιβεβαιώνεται μέσω ενός βοηθητικού παραθύρου που τοποθετείται γύρω από αυτό και στο οποίο ψάχνουμε με μικρότερο βήμα για πιο ακριβή εντοπισμό. Εάν επιβεβαιωθεί, η όλη διαδικασία ξεκινάει από το τελικό αποτέλεσμα, διαφορετικά γίνεται το ίδιο ξεκινώντας από το απορριφθέν σημείο

αλλαγής. Η σχηματική περιγραφή του αλγορίθμου δίνεται στο Σχήμα 3.4.



Σχήμα 3.4

Τέλος, μένει η ένωση των τμημάτων που ανήκουν στον ίδιο ομιλητή. Αυτό γίνεται με τον ίδιο ακριβώς τρόπο, συγκρίνοντας "απο κάτω προς τα πάνω" τα απομονωμένα τμήματα και ενώνοτας διαδοχικά αυτά τα οποία απέχουν λιγότερο.

Ένα χαρακτηριστικό της παραπάνω μεθόδου, είναι ότι σε κάθε στιγμή, βρίσκεται υπό επεξεργασία μόνο ένα τμήμα του σήματος ομιλίας. Αυτή είναι μία πολύ χρήσιμη ιδιότητα, ειδικά αν ληφθούν υπόψη εφαρμογές όπου απαιτείται επεξεργασία σε πραγματικό χρόνο. Το πιο σημαντικό όμως, είναι ότι η μέθοδος αποτελείται από δύο *ανεξάρτητα* κομμάτια: από έναν τρόπο αναζήτησης ενός σημείου αλλαγής και από έναν τρόπο σύγκρισης δύο γειτονικών τμημάτων ομιλίας. Στη βιβλιογραφία υπάρχει πλήθος εναλλακτικών προσεγγίσεων όπου αλλάζει το ένα ή και τα δύο από αυτά, πάντα όμως στην ίδια φιλοσοφία. Για παράδειγμα, η απόφαση για ύπαρξη σημείου αλλαγής μπορεί να γίνει χρησιμοποιώντας κάποιο διαφορετικό μέτρο όπως τα

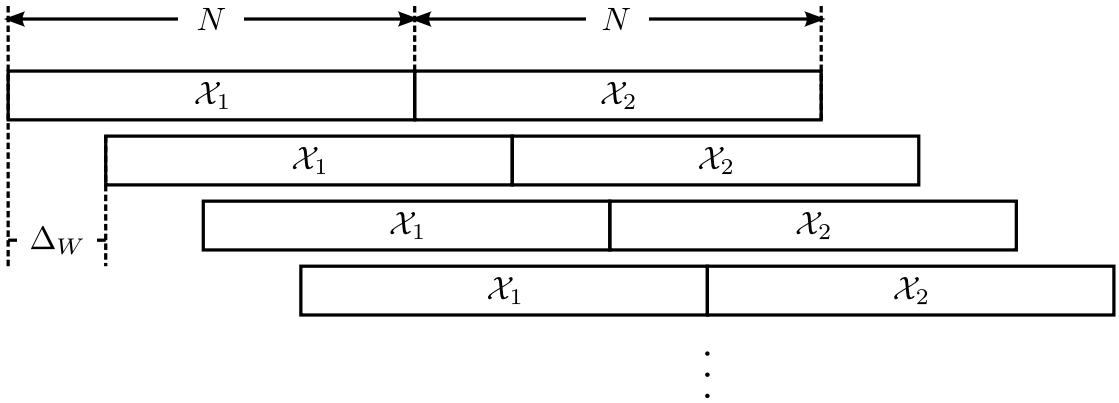
συμμετρική απόσταση Kullback-Leibler

$$\begin{aligned} KL2(i) =& \frac{1}{2}(\mathbf{m}_1 - \mathbf{m}_2)^T \left(S_1^{-1} + S_2^{-1}\right)(\mathbf{m}_1 - \mathbf{m}_2) \\ &+ \frac{1}{2} tr\left(S_1^{-1}S_2 + S_2^{-1}S_1 - 2I\right) \end{aligned} \tag{3.11}$$

κριτήριο T^2 Hotelling

$$T^2(i) = \frac{i(N-i)}{N}(\mathbf{m}_1 - \mathbf{m}_2)^T S^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \tag{3.12}$$

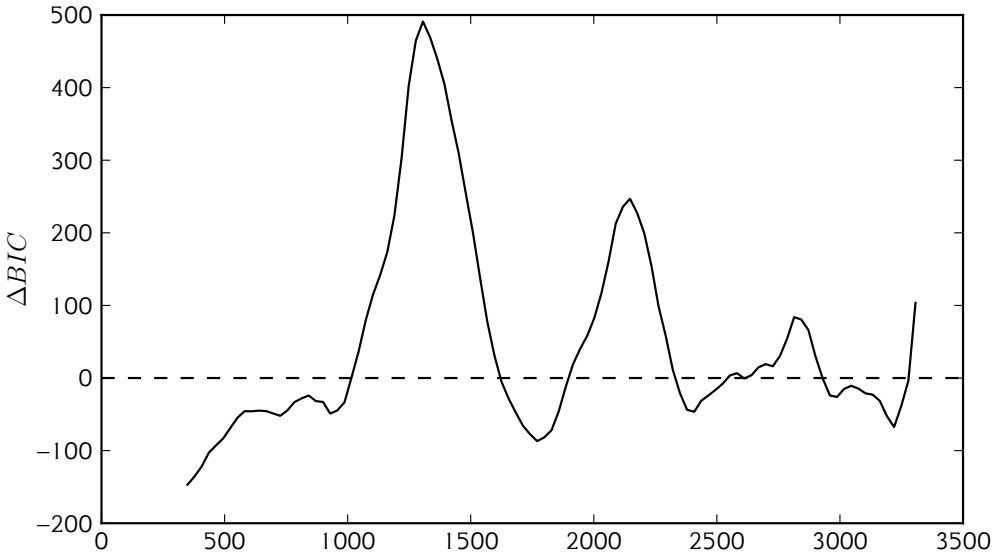
Επίσης, η αναζήτηση του σημείου αλλαγής μπορεί να γίνει με διαφορετικό τρόπο. Μία πολύ απλή προσέγγιση είναι αυτή του Σχήματος 3.5 [Cheng et al., 2010].



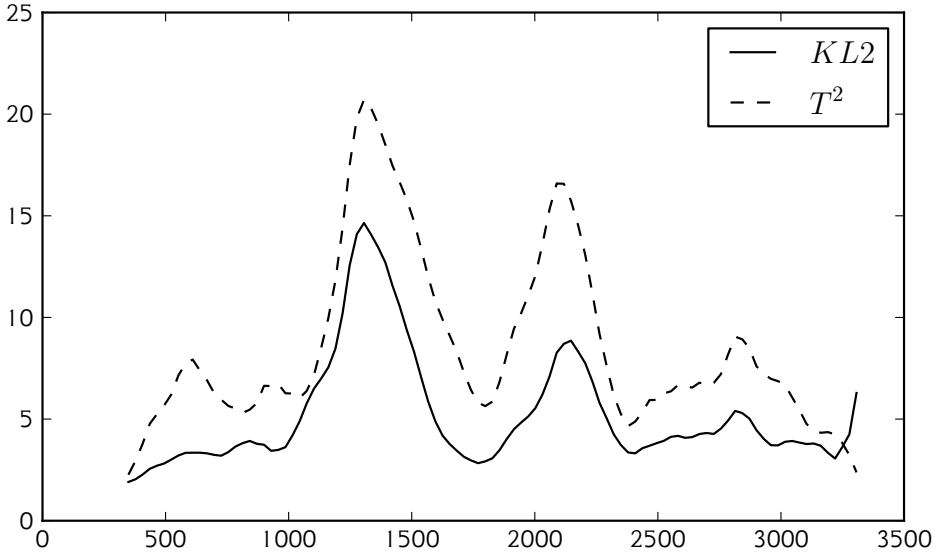
Σχήμα 3.5

Εδώ, υπάρχουν δύο ισομήκη και εφαπτόμενα παράθυρα, τα οποία μετατοπίζονται με ένα σταθερό βήμα. Σε κάθε τέτοιο βήμα υπολογίζεται η "απόσταση" των δύο τμημάτων και το σημείο αλλαγής προσδιορίζεται από τις μέγιστες τιμές αυτής της απόστασης. Μία εικόνα των αποτελεσμάτων που παίρνουμε από αυτή τη μέθοδο για διάφορα μέτρα δίνεται στα Σχήματα 3.6 και 3.7. Το σήμα που έχει χρησιμοποιηθεί περιέχει τρία σημεία αλλαγής που ξεχωρίζουν από τις κορυφές των γραφημάτων.

Φαίνεται ότι η επιλογή μέτρου δεν είναι τόσο σημαντική αφού τα αποτελέσματα παρουσιάζουν μεγάλη ομοιότητα. Εδώ όμως βρίσκεται και η μεγαλύτερη αδυναμία αυτής της οικογένειας μεθόδων: Όπως με το μέτρο ΔBIC, είναι αναγκαία μία κατωφλική τιμή για την απόφαση αποδοχής ή απόρριψης μιας κορυφής του γραφήματος



Σχήμα 3.6



Σχήμα 3.7

ως σημείο αλλαγής ομιλητή. Τέτοιες τιμές μπορούν να εξαχθούν θεωρητικά, υπό την προϋπόθεση ότι γνωρίζουμε την κατανομή των παρατηρήσεων (θυμίζουμε ότι σε αυτή

την υπόθεση στηρίζεται και η κατωφλική τιμή του μέτρου ΔBIC). Στην πράξη όμως κάτι τέτοιο δεν μπορεί να γίνει και οι θεωρητικές τιμές δεν μπορούν να χρησιμοποιηθούν. Ακόμα και στην εξίσωση (3.5) υπάρχει ένας συντελεστής ποινής λ που παίζει αυτόν το ρόλο. Έτσι, αν και στη βιβλιογραφία έχουν προταθεί τρόποι για να μειωθεί η ανάγκη εμπειρικού προσδιορισμού της κατωφλικής τιμής (π.χ [Ajmera et al., 2004]), αυτός πρακτικά γίνεται είτε με ευρετικές μεθόδους, είτε μέσω εκπαίδευσης, όπου επιλέγεται η τιμή που δίνει τα καλύτερα αποτελέσματα σε σχέση με κάποια ήδη διαχωρισμένα σήματα ομιλίας,

3.2 Η μέθοδος FLsD

Ένα θέμα που δεν θίχτηκε στην προηγούμενη παράγραφο είναι αυτό της διάστασης των χαρακτηριστικών η οποία συχνά είναι αρκετά μεγάλη, ειδικά όταν γίνεται συνδυασμός χαρακτηριστικών. Η αποδοτικότητα των διαφόρων μεθόδων εξαρτάται με δύο τρόπους από τη διάσταση των χαρακτηριστικών, αφού αυτή πρέπει να είναι *αρκετά μεγάλη* ώστε να περιέχεται όλη η πληροφορία που αφορά το διαχωρισμό ομιλητών, αλλά συγχρόνως *όχι υπερβολικά μεγάλη* ωστε να περιλαμβάνονται άσχετα ως προς το διαχωρισμό χαρακτηριστικά. Μία πρόταση προς αυτή την κατεύθυνση δίνεται στα [Kotti et al., 2006] και [Kotti et al., 2008]. Εδώ, χρησιμοποιείται ένας μεγάλος αριθμός χαρακτηριστικών και ο προσδιορισμός του βέλτιστου συνδυασμού τους που τελικά χρησιμοποιείται γίνεται με χρήση μιας βιβλιοθήκης προ-διαχωρισμένων ηχογραφήσεων ομιλίας και μεθόδους διακλάδωσης και οριοθέτησης [van der Heijden et al., 2004]. Μια παρόμοια προσέγγιση γίνεται στο [Chu et al., 2009], όπου πάλι με χρήση βιβλιοθήκης προ-διαχωρισμένων ηχογραφήσεων, προσδιορίζεται ο βέλτιστος υπόχωρος μέσω γραμμικής διαχωριστικής ανάλυσης θεωρώντας κάθε ομιλητή ως μία κλάση. Στόχος αυτών των ημι-επιβλεπόμενων μεθόδων είναι η εύρεση ενός υπόχωρου που θεωρείται *γενικά* βέλτιστος για τους σκοπούς του διαχωρισμού ομιλητών, ο υπολογισμός του οποίου όμως, βασίζεται αποκλειστικά στις διαθέσιμες ηχογραφήσεις. Αυτός είναι και ο κύριος λόγος που είναι επιθυμητό να χρησιμοποιούνται μη-επιβλεπόμενες μέθοδοι (πέρα από τη χρονοβόρα διαδικασία της εκπαίδευσης), όπως αυτή που παρουσιάζεται στο [Castaldo et al., 2008], όπου η μείωση διαστάσεων γίνεται μέσω

ανάλυσης κύριων συνιστωσών. Τα προβλήματα αυτής της προσέγγισης είναι αυτά που συζητήθηκαν και στο προηγούμενο κεφάλαιο, δηλαδή ότι δεν γνωρίζουμε εάν ο υπόχωρος που δημιουργείται είναι κατάλληλος για διαχωρισμό ομιλητών.

Η μέθοδος που προτείνεται στο [Giannakopoulos and Petridis, 2012] επιχειρεί να λύσει και τα δύο προβλήματα ταυτόχρονα, κατασκευάζοντας έναν σχεδόν βέλτιστο για διαχωρισμό υπόχωρο, αποκλειστικά για το υπό επεξεργασία σήμα και χωρίς να προηγείται κάποιο στάδιο εκπαίδευσης. Η κεντρική ιδέα, είναι ότι ακόμα και αν δεν έχουμε έτοιμη την πληροφορία για το ποιος μιλάει, μπορούμε να εξάγουμε κάποια σχετική πληροφορία από την ίδια τη φύση της ομιλίας. Πιο συγκεκριμένα, μπορούμε να υποθέσουμε ότι σε κάθε μικρό τμήμα ομιλίας π.χ. τάξης ενός δευτερολέπτου, μιλάει *ένας* (άγνωστος) ομιλητής. Αν κάτι τέτοιο ισχύει, δεν ξέρουμε μεν σε ποια κλάση ανήκει αυτό το τμήμα ομιλίας, αλλά ξέρουμε ότι όλα τα διανύσματα χαρακτηριστικών που περιέχονται σε αυτό, προέρχονται από την ίδια κλάση. Χρησιμοποιούμε τον όρο *νήμα κλάσης* για να περιγράψουμε ένα σύνολο παρατηρήσεων που ανήκουν στην ίδια κλάση. Το πρόβλημα του διαχωρισμού ομιλητών με αυτό τον τρόπο να ανάγεται στον προσδιορισμό ενός συνόλου αντιστοιχήσεων από νήματα κλάσης σε κλάσεις. Στη μέθοδο γραμμικής ημι-διαχωριστικής ανάλυσης Fisher (**F**isher **L**inear **s**emi-**D**iscriminant Analysis - FLsD), εκτιμάται ότι ο υπόχωρος που κατασκευάζεται με τεχνικές της γραμμικής διαχωριστικής ανάλυσης παίρνοντας σαν κλάσεις τα νήματα κλάσης, είναι αρκετά κοντά με αυτόν που θα παίρναμε εάν γνωρίζαμε τις κλάσεις εκ των προτέρων. Αυτά θα γίνουν πιο καθαρά στη συνέχεια, όπου θα παρουσιάσουμε τις ιδιότητες των νημάτων κλάσης και τη σχέση τους με αυτές των πραγματικών κλάσεων, χτίζοντας έτσι το θεωρητικό υπόβαθρο της μεθόδου.

3.2.1 Θεωρία

Ξεκινάμε δίνοντας τον ορισμό του χώρου όπου θα εργαστούμε:

Ορισμός 2 (Γραμμικός ημι-διαχωριστικός υπόχωρος Fisher)**.** *Έστω ένα σύνολο παρατηρήσεων* $\mathcal{X} = \{\mathbf{x}_i\} \subset \mathbb{R}^d$*, ένα σύνολο κλάσεων* $\mathcal{C} = \{c_k\}$ *και το σύνολο των αντιστοιχίσεων κάθε παρατήρησης σε μία κλάση* $\{\mathbf{x}_i, c^i\}$*. Έστω επίσης ένα σύνολο νημάτων κλάσης*

$\mathcal{V}$, $|\mathcal{V}| > |\mathcal{C}|$, η επιρριπτική αντιστοίχιση $h : \mathcal{V} \to \mathcal{C}$ που αντιστοιχίζει κάθε νήμα κλάσης σε μία αρχική κλάση και το σύνολο αντιστοιχίσεων των παρατηρήσεων σε νήματα κλάσης $\{\mathbf{x}_i, v^i\}$ τέτοια ώστε $\forall i : h(v^i) = c^i$. Συμβολίζουμε με S_w^h και S_b^h τους πίνακες διασποράς εντός νημάτων κλάσης και μεταξύ νημάτων κλάσης αντίστοιχα και με S_m τον πίνακα ολικής διασποράς. Για κάθε $d' < d$, ο διάστασης d' υπόχωρος του R^d που δημιουργείται από τις στήλες της λύσης του προβλήματος βελτιστοποίησης

$$\hat{A} = \underset{A \in \mathbb{R}^{d \times d'}}{\operatorname{argmax}}\, r(A, S_1, S_2)$$

όπου (S_1, S_2) οποιοδήποτε από τα $\left(S_m, S_w^h\right)$, $\left(S_b^h, S_w^h\right)$, $\left(S_b^h, S_m\right)$ και r κάποια από τις συναρτήσεις (2.47), (2.50), ορίζεται ως βέλτιστος γραμμικός ημι-διαχωριστικός υπόχωρος Fisher.

Ορισμός 3 (Πίνακας διαταραχής νήματος κλάσης)**.** *Έστω ένα νήμα κλάσης v και $\mathbf{d}_v$ η διαφορά του δειγματικού μέσου της αντίστοιχης κλάσης $h(v)$ από το δειγματικό μέσο του νήματος κλάσης v*

$$\mathbf{d}_v = \mathbf{m}_{h(v)} - \mathbf{m}_v \tag{3.13}$$

Ο πίνακας διαταραχής νήματος κλάσης S_h ορίζεται ως ο πίνακας συνδιακύμανσης αυτών των διαφορών για όλα τα νήματα κλάσης

$$S_h = \underset{v \in \mathcal{V}}{\mathcal{E}} \left[\mathbf{d}_v \mathbf{d}_v^T\right] \tag{3.14}$$

Ο όρος "πίνακας διαταραχής" δικαιολογείται από το ακόλουθο λήμμα

Λήμμα 3.2.0.1. *Για κάθε αντιστοίχηση νημάτων κλάσης h ισχύουν*

$$S_w^h = S_w + S_h, \text{ και} \tag{3.15α'}$$

$$S_b^h = S_b - S_h \tag{3.15β'}$$

Απόδειξη. Παίρνουμε τα δείγματα ενός συγκεκριμένου νήματος κλάσης v. Ο πίνακας συνδιακύμανσής τους είναι

$$\underset{\mathbf{x} \mapsto v}{\mathcal{E}} \left[\mathbf{x}\mathbf{x}^T\right] - \mathbf{m}_v \mathbf{m}_v^T \tag{3.16}$$

ενώ ο πίνακας συνδιακύμανσης για τα ίδια δείγματα υπό την υπόθεση ότι ο μέσος της αντίστοιχης κλάσης είναι γνωστός, είναι

$$\underset{\mathbf{x} \mapsto v}{\mathcal{E}} \left[\mathbf{x}\mathbf{x}^T\right] - \mathbf{m}_{h(v)} \mathbf{m}_{h(v)}^T \tag{3.17}$$

Παίρνοντας τη διαφορά τους και το μέσο όρο για όλες τις κλάσεις, έχουμε

$$S_h = S_w - S_w^h = \underset{v\in\mathcal{V}}{\mathcal{E}}\left[\mathbf{m}_v\mathbf{m}_v^T - \mathbf{m}_{h(v)}\mathbf{m}_{h(v)}^T\right] \tag{3.18}$$

Θέτουμε $\mathbf{d}_v = \mathbf{m}_{h(v)} - m_v$ και η παραπάνω διαφορά γράφεται ως

$$\begin{aligned} S_h &= \underset{v\in\mathcal{V}}{\mathcal{E}}\left[\left(\mathbf{m}_{h(v)} - \mathbf{d}_v\right)\left(\mathbf{m}_{h(v)} - \mathbf{d}_v\right)^T - \mathbf{m}_{h(v)}\mathbf{m}_{h(v)}^T\right] \\ &= \underset{v\in\mathcal{V}}{\mathcal{E}}\left[\mathbf{d}_v\mathbf{d}_v^T - \mathbf{m}_{h(v)}\mathbf{d}_v^T - \mathbf{d}_v\mathbf{m}_{h(v)}^T\right] \\ &= \underset{v\in\mathcal{V}}{\mathcal{E}}\left[\mathbf{d}_v\mathbf{d}_v^T\right] + \underset{c\in\mathcal{C}}{\mathcal{E}}\left[\mathbf{m}_c\left(\underset{v:h(v)=c}{\mathcal{E}}\left[\mathbf{d}_v\right]\right)^T + \mathbf{m}_c^T\left(\underset{v:h(v)=c}{\mathcal{E}}\left[\mathbf{d}_v\right]\right)\right] \end{aligned} \tag{3.19}$$

Εξ' ορισμού, για όλες τις κλάσεις c έχουμε

$$\forall c \in \mathcal{C} : \quad \underset{v:h(v)=c}{\mathcal{E}}\left[\mathbf{d}_v\right] = \mathbf{0} \tag{3.20}$$

και συνεπώς ο δεύτερος όρος της εξίσωσης (3.19) μηδενίζεται, άρα

$$S_h = \underset{v\in\mathcal{V}}{\mathcal{E}}\left[\mathbf{d}_v\mathbf{d}_v^T\right] \tag{3.21}$$

□

Λήμμα 3.2.0.2. *Έστω ότι* $\forall v \in \mathcal{V} : \left\|\mathbf{m}_{h(v)} - \mathbf{m}_v\right\|_2 \leq \delta$ *για κάποιο* $\delta > 0$. *Τότε*

$$\left\|S_h\right\|_2 \leq \delta^2$$

Απόδειξη. Από τον ορισμό του πίνακα S_h, έχουμε

$$\left\|S_h\right\|_2 = \left\|\underset{v}{\mathcal{E}}\left[\mathbf{d}_v\mathbf{d}_v^T\right]\right\|_2 \leq \underset{v}{\mathcal{E}}\left[\left\|\mathbf{d}_v\mathbf{d}_v^T\right\|_2\right] = \underset{v}{\mathcal{E}}\left[\left\|\mathbf{d}_v\right\|_2^2\right] \leq \delta^2$$

□

Η αποτελεσματικότητα της χρήσης νημάτων κλάσης μπορεί να υποστηριχθεί θεωρητικά από αποτελέσματα της θεωρίας διαταραχών πινάκων. Δεν θα παραθέσουμε αυτά τα αποτελέσματα στην πληρότητά τους, ούτε τις μακροσκελείς αποδείξεις που τα συνοδεύουν, καθώς η σχετική βιβλιογραφία είναι εκτενής και πλήρης. Θα παραθέσουμε όμως μερικά αποτελέσματα της θεωρίας που δίνουν μια ιδέα για το πως προσεγγίζεται ο υπόχωρος LDA από τα νήματα κλάσης.

Το πρώτο θεώρημα [Li, 1996] δίνει ένα άνω φράγμα για τη διαφορά των ιδιοτιμών και το δεύτερο δείχνει πως αυτό το άνω φράγμα μπορεί να προσαρμοστεί σε *γενικευμένα* προβλήματα ιδιοτιμών.

Θεώρημα 3.2.1 (Lidskii-Mirsky-Wielandt). *Έστω πίνακας A με ιδιοτιμές $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$ και μια διαταραχή του $A + \Delta A$ με ιδιοτιμές $\tilde{\lambda}_1 \leq \tilde{\lambda}_2 \leq \cdots \leq \tilde{\lambda}_d$. Τότε ισχύει*

$$|\lambda_i - \tilde{\lambda}_i| \leq \|\Delta A\|_2, \quad \forall i = 1, 2, \ldots, d$$

Θεώρημα 3.2.2. *Τότε ισχύει*

$$|\lambda_i - \tilde{\lambda}_i| \leq \frac{\Delta A}{\lambda_{min}(B)}$$

Απόδειξη. Θεωρούμε τον πίνακα $\mathrm{Z} = \mathrm{B}^{-1/2}$. Εύκολα βλέπουμε ότι

$$A\mathbf{x} = \lambda B\mathbf{x} \Rightarrow (ZAZ)\mathbf{x} = \lambda(ZBZ)\mathbf{x} = \lambda\mathbf{x}$$

δηλαδή οι ιδιοτιμές του ZAZ είναι ίδιες με τις γενικευμένες ιδιοτιμές της δέσμης (A, B) □

Παρόμοια αποτελέσματα δίνονται και για την απόσταση του υπόχωρου που δημιουργείται από τα γενικευμένα ιδιοδιανύσματα της δέσμης (S_w^h, S_m) και του βέλτιστου υπόχωρου που αποκτάται μέσω γραμμικής διαχωριστικής ανάλυσης μέσω θεωρημάτων όπως το παρακάτω [Li, 1996]:

Θεώρημα 3.2.3 (Davis-Kahan). *Έστω ένας ερμιτιανός πίνακας A και μία διαταραχή του $\bar{A}$ με παραγοντοποιήσεις ιδιοτιμών*

$$A = U\Lambda U^* = \left(U_1, U_2\right)\begin{pmatrix} \Lambda_1 & \\ & \Lambda_2 \end{pmatrix}\begin{pmatrix} U_1^* \\ U_2^* \end{pmatrix}$$
$$\bar{A} = \bar{U}\Lambda\bar{U}^* = \left(\bar{U}_1, \bar{U}_2\right)\begin{pmatrix} \bar{\Lambda}_1 & \\ & \bar{\Lambda}_2 \end{pmatrix}\begin{pmatrix} \bar{U}_1^* \\ \bar{U}_2^* \end{pmatrix}$$

όπου $U, \bar{U} \in \mathbb{R}^{d\times d}$, $U_1, \bar{U}_1 \in \mathbb{R}^{d\times k}$ και

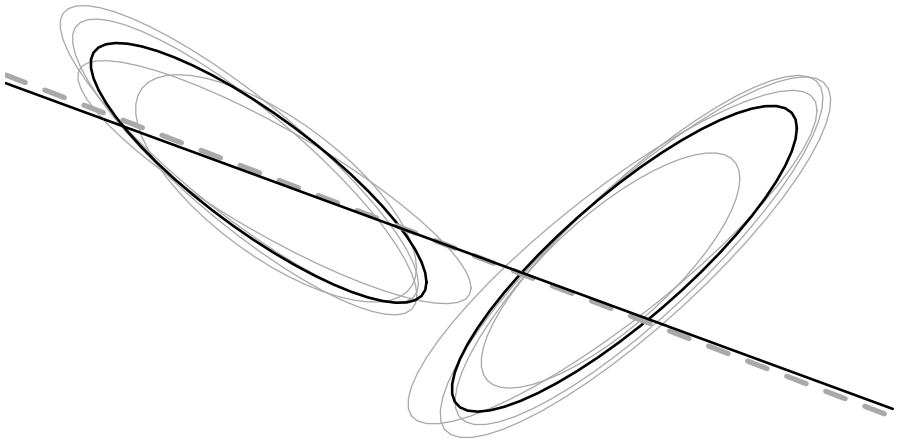
$$\Lambda_1 = diag(\lambda_1, \ldots, \lambda_k), \ \Lambda_2 = diag(\lambda_{k+1}, \ldots, \lambda_d)$$
$$\bar{\Lambda}_1 = diag(\bar{\lambda}_1, \ldots, \bar{\lambda}_k), \ \bar{\Lambda}_2 = diag(\bar{\lambda}_{k+1}, \ldots, \bar{\lambda}_d)$$

Τότε, εάν $\delta := \min_{1\leq i\leq k,\, 1\leq j\leq d-k} |\lambda_i - \bar{\lambda}_{k+j}| > 0$, ισχύει

$$\left\|\sin\Theta(U_1, \bar{U}_1)\right\|_F \leq \frac{\left\|(\bar{A} - A)U_1\right\|_F}{\delta}$$

Με παρόμοιο σκεπτικό μπορούν να εξαχθούν άνω φράγματα για όλες τις δέσμες πινάκων που προτείνονται και για το κριτήριο (2.50).

Σε κάθε περίπτωση, καταλήγουμε στο συμπέρασμα ότι μπορούμε να περιμένουμε αποτελέσματα αρκετά κοντά σε αυτά που θα παίρναμε μέσω γραμμικής διαχωριστικής ανάλυσης, αρκεί οι μέσοι των νημάτων κλάσης να μην απέχουν πολύ από αυτούς της αντίστοιχης κλάσης (Σχήμα 3.8).



Σχήμα 3.8

3.2.2 Εφαρμογή

Περιγράφουμε εδώ τον τρόπο με τον οποίο μπορεί να εφαρμοστεί η παραπάνω θεωρία στα πλαίσια του διαχωρισμού ομιλητών και συγκεκριμένα τη μεθοδολογία που ακολουθείται στο [Giannakopoulos and Petridis, 2012] αλλά και στο λογισμικό diarx που αποτελεί μέρος της παρούσας εργασίας. Όπως αναφέραμε στην αρχή του κεφαλαίου, στόχος είναι η δημιουργία ενός υπόχωρου που διαχωρίζει τους ομιλητές του σήματος, οπότε υποθέτουμε ότι διαθέτουμε ολόκληρο το σήμα ομιλίας διάρκειας T (σε αντίθεση με τις μεθόδους της προηγούμενης ενότητας που αρκούσε ένα τμήμα του κάθε φορά).

Αρχικά, στα πλαίσια βραχύχρονης επεξεργασίας του σήματος λαμβάνονται $d/2$

συντελεστές MFCC από πλαίσια μήκους w_s

$$\left\{\tilde{\mathbf{x}}[n] \in \mathbb{R}^{d/2}\right\},\ n = 1, 2, \ldots, T/w_s$$

Για κάθε L συνεχόμενα διανύσματα MFCC που εξάγονται, υπολογίζονται οι μέσοι όροι και οι διασπορές των συντελεστών τους. Συγκεκριμένα έχουμε

$$x_i[n] = \begin{cases} \frac{1}{L}\sum_{m=n}^{n+L} \tilde{x}_i[m] & i = 1, \ldots, d/2 \\ \frac{1}{L}\sum_{m=n}^{n+L} \left(\tilde{x}_{i-d/2}[m] - x_{i-d/2}[n]\right)^2 & i = d/2+1, \ldots, d \end{cases} \tag{3.22}$$

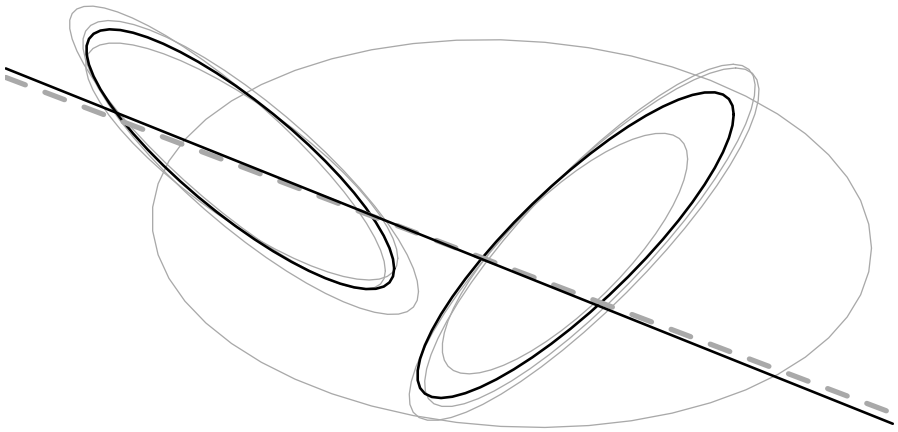
Κάθε διάνυσμα $\mathbf{x}[n]$ αντιστοιχεί σε χρόνο $L \cdot w_s$ και μπορεί να θεωρηθεί σαν ένα χαρακτηριστικό μέσου χρόνου (ή κατά την ορολογία του [Tzanetakis and Cook, 2002] όπου ακολουθείται παρόμοια μεθοδολογία, ένα *παράθυρο υφής*).

Τέλος, για τη δημιουργία των νημάτων κλάσης θεωρούμε ένα μήκος παραθύρου w_l όπου θέλουμε

1. σε κάθε τμήμα σήματος διάρκειας w_l υπάρχει μόνο ένας ομιλητής
2. οι μέσοι όροι των διανυσμάτων που περιέχονται σε κάθε παράθυρο μήκους w_l δεν πρέπει να απέχουν πολύ από τους μέσους όρους των διανυσμάτων που αντιστοιχούν στον ομιλητή του παραθύρου.

Οι δύο υποθέσεις που πρέπει να ικανοποιεί αυτό το παράθυρο είναι αντιφατικές, αφού από τη μία το παράθυρο πρέπει να είναι αρκετά μικρό ώστε να ικανοποιείται η πρώτη, αλλά και αρκετά μεγάλο ώστε να μην παραβιάζεται η δεύτερη εξαιτίας του μικρού δείγματος. Μία ρεαλιστική τιμή που δίνει ικανοποιητικά αποτελέσματα είναι $w_l \simeq 1$ sec.

Σημείωση 4. *Η απαίτηση του να υπάρχει μόνο ένας ομιλητής σε κάθε παράθυρο μήκους w_l μπορεί να ισχύσει μόνο κατά προσέγγιση, αφού σε κάποια από αυτά θα υπάρχουν αλλαγές ομιλητή. Ένας τρόπος να αποφευχθεί αυτό, είναι να γίνεται έλεγχος για τέτοιες περιπτώσεις π.χ. μέσω του κριτηρίου ΔBIC και αν παρατηρούνται αλλαγές αυτό το παράθυρο απλά να αγνοείται. Στην πράξη πάντως, αυτά τα παράθυρα αποτελούν ένα πολύ μικρό μέρος του συνόλου και δεν προκαλούν προβλήματα στην απόδοση της μεθόδου. Σαν παράδειγμα μπορούμε να δούμε το Σχήμα 3.9 σε αντιπαραβολή με το*



Σχήμα 3.9

Σχήμα 3.8. Βλέπουμε ότι το ένα νήμα κλάσης περιέχει δείγματα και από τις δύο αρχικές κλάσεις, αλλά η ευθεία προβολής δεν αλλάζει σημαντικά.

Είμαστε τώρα σε θέση να παρουσιάσουμε τον αλγόριθμο υπολογισμού του (σχεδόν) βέλτιστου υπόχωρου για διαχωρισμό ομιλητών

Αλγόριθμος 6 Υπολογισμός πίνακα μετασχηματισμού FLsD

Είσοδος: Τα χαρακτηριστικά μέσου χρόνου $\mathbf{x}[n],\ n = 1, 2, \ldots, N$

Έξοδος: Ο $d \times d'$ πίνακας προβολής $\hat{A}$

1: Αρχικοποίηση δεικτών ακολουθίας παραθύρων και νήματος κλάσης

$$n \leftarrow 1, \quad \nu \leftarrow 1$$

2: Αρχικοποίηση στατιστικών μεγεθών

$$\mathbf{m} = \mathbf{0} \in \mathbb{R}^d, \quad S_m \leftarrow 0 \in \mathbb{R}^{d\times d}, \quad S_w^h \leftarrow 0 \in \mathbb{R}^{d\times d}$$

3: **για όσο** $n \leq N - L$

4:

$$R \rightarrow [n, \ldots, n + L]$$

5: Υπολογισμός στατιστικών νήματος κλάσης

$$\mathbf{m}_c \leftarrow \frac{1}{|R|}\sum_{n\in R}\mathbf{x}[n], \quad S_c \leftarrow \frac{1}{|R|}\sum_{n\in R}\mathbf{x}[n]\mathbf{x}[n]^T$$

6: Ανανέωση πίνακα διασποράς εντός κλάσεων

$$S_w^h \leftarrow S_w^h + \frac{w_l}{T}\left(S_c - \mathbf{m}_c\mathbf{m}_c^T\right)$$

7: Ανανέωση συνολικών στατιστικών μεγεθών

$$\mathbf{m} \leftarrow \mathbf{m} + \frac{w_l}{T}\mathbf{m}_c, \quad S_m \leftarrow \frac{w_l}{T}S_c$$

8: Αύξηση τιμής δεικτών

$$\nu \leftarrow \nu + 1, \quad n \leftarrow n + L$$

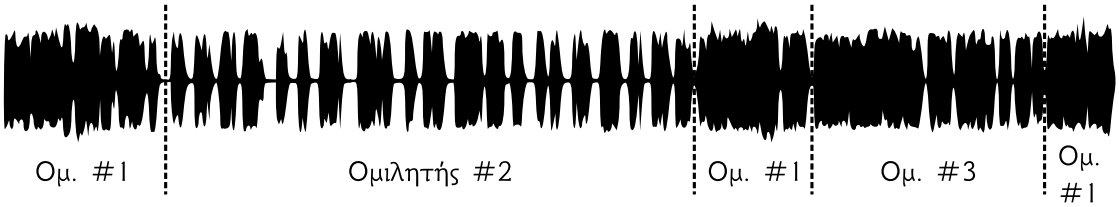
9: $S_m \leftarrow S_m = \mathbf{m}\mathbf{m}^T$

10: $\hat{A} = \operatorname{argmax}_{A\in\mathbb{R}^{d\times d'}} r(A, S_m, S_w^h)$

Η όλη διαδικασία ολοκληρώνεται με ομαδοποίηση των χαρακτηριστικών μειωμένης διάστασης χρησιμοποιώντας τη μέθοδο ασαφών K-μέσων.

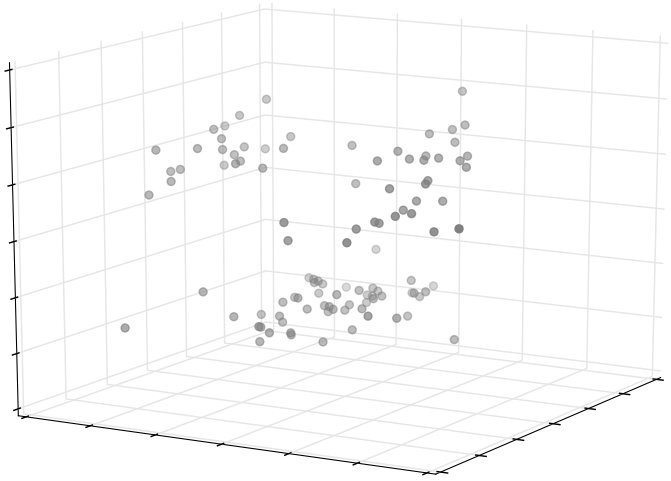
Σημείωση 5. *Η επιλογή των συντελεστών MFCC ως βάση των χαρακτηριστικών που χρησιμοποιούνται, είναι απλά μια λεπτομέρεια της υλοποίησης της μεθόδου. Αντί αυτών μπορεί να γίνει χρήση κάθε χαρακτηριστικού που διευκολύνει το διαχωρισμό ομιλητών (π.χ. συντελεστές LSP) ή και συνδυασμός τους.*

Αφήνουμε για το τέλος ένα παράδειγμα εφαρμογής της μεθόδου. Το σήμα του Σχήματος 3.10 είναι ένα απόσπασμα από κεντρικό δελτίο ειδήσεων. Υπάρχουν συνολικά τρεις ομιλητες, η παρουσιάστρια του δελτίου (1), ένας δημοσιογράφος (2) και ένας πολιτικός (3) και η σειρά με την οποία μιλούν είναι: (1) - (2) - (1) - (3) - (1).



Σχήμα 3.10

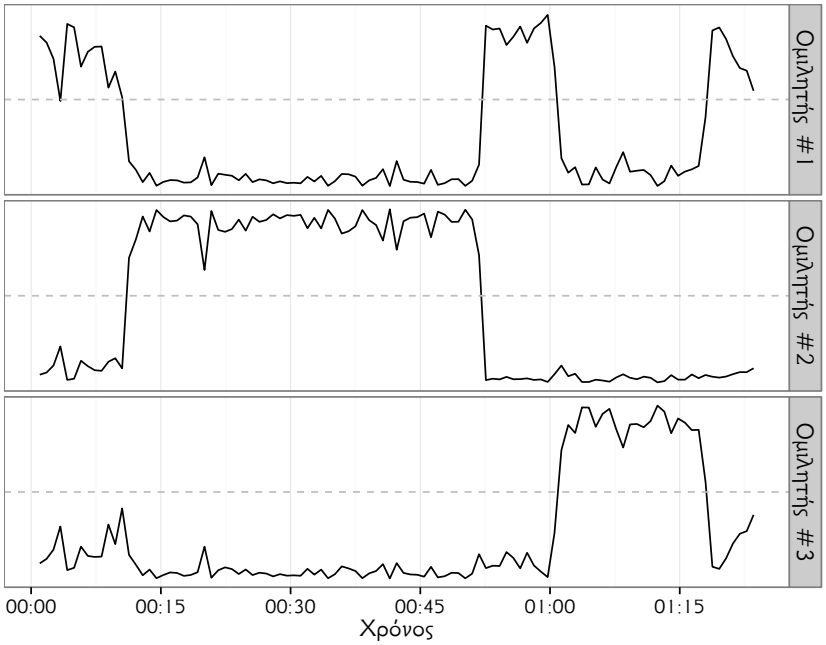
Στο Σχήμα 3.11 βλέπουμε την προβολή των χαρακτηριστικών που εξάχθηκαν από το σήμα, σε υπόχωρο τριών διάστασεων μέσω του Αλγόριθμου 6. Ιδανικά θα έπρεπε



Σχήμα 3.11

να έχουμε τέσσερις διαστάσεις για διαχωρισμό τριών ομιλητών, αλλά η επιλογή έγινε για λόγους οπτικοποίησης των αποτελεσμάτων. Σε κάθε περίπτωση όμως, οι τρεις ομάδες που περιμέναμε σχηματίζονται και είναι ευδιάκριτες.

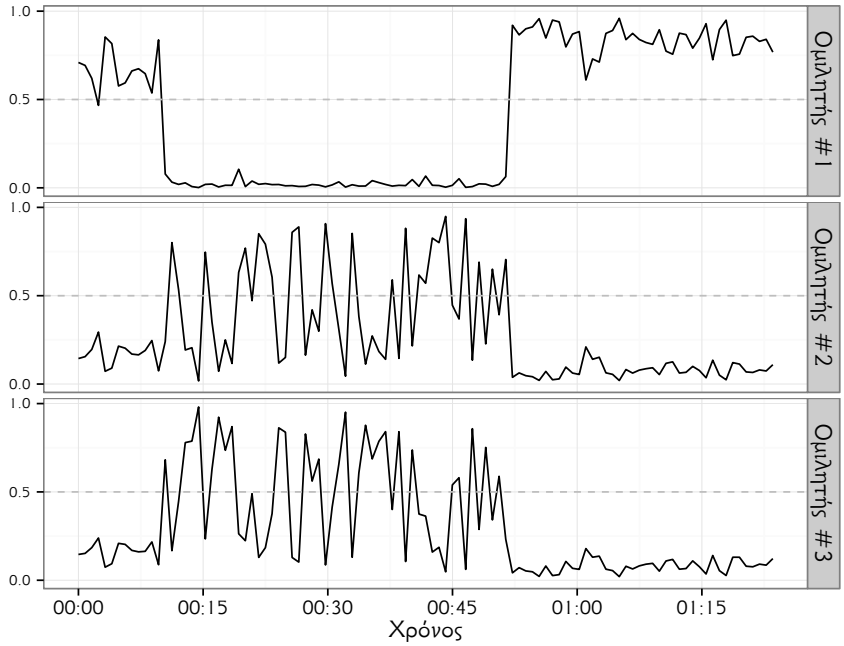
Το αποτέλεσμα της ασαφούς ομαδοποίησής τους φαίνεται στο Σχήμα 3.12.



Σχήμα 3.12

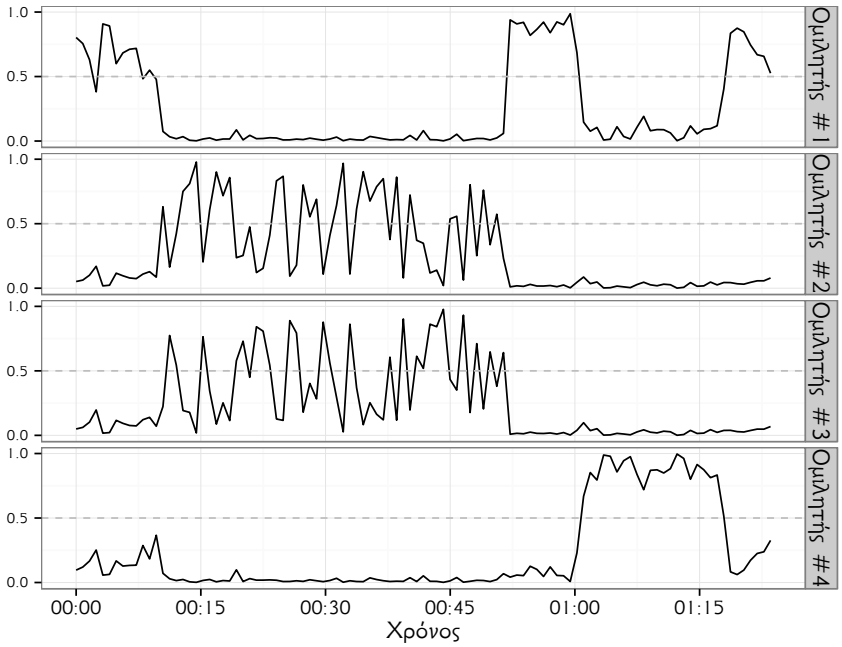
Αυτό που έχει σχεδιαστεί είναι το ποσό συμμετοχής του κάθε δείγματος σε κάθε ομάδα, ή ισοδύναμα, κάθε γραμμή του σχήματος αντιστοιχεί σε μία γραμμή του πίνακα συμμετοχής της ασαφούς ομαδοποίησης. Ταυτόχρονα όμως, αυτό μπορεί να ερμηνευθεί και ως *πιθανότητα το δείγμα να ανήκει στον κάθε ομιλητή*, συνεπώς η απόφαση του "ποιος μιλάει" γίνεται με βάση τη μεγαλύτερη πιθανότητα.

Ο διαχωρισμός ομιλητών στο παραπάνω σχήμα είναι επιτυχής. Υπάρχουν όμως δύο σημαντικοί παράγοντες που μπορούν να οδηγήσουν σε αποτυχία του. Ο πρώτος είναι η λανθασμένη ομαδοποίηση. Αναφέραμε και στο προηγούμενο κεφάλαιο ότι οι επαναληπτικές μέθοδοι που παρουσιάσαμε (και χρησιμοποιούμε εδώ), είναι δυνατόν να συγκλίνουν σε ένα τοπικό ελάχιστο της αντικειμενικής τους συνάρτησης εξαιτίας κακής αρχικοποίησης. Μία τέτοια περίπτωση φαίνεται στο Σχήμα 3.13.



Σχήμα 3.13: Διαχωρισμός με ανεπιτυχή ομαδοποίηση ομιλητών

Ο δεύτερος και ίσως πιο σημαντικός παράγοντας, είναι το ότι στην πράξη σπάνια γνωρίζουμε τον αριθμό ομιλητών που περιέχονται στο σήμα εκ των προτέρων. Στο Σχήμα 3.14 βλέπουμε το αποτέλεσμα του διαχωρισμού με υπόθεση ότι υπάρχουν τέσσερις ομιλητές.



Σχήμα 3.14: Διαχωρισμός με λάθος εκτίμηση αριθμού ομιλητών

Η εκτίμηση του αριθμού ομιλητών μπορεί να γίνει με διάφορους τρόπους. Ένας από αυτούς είναι και η οπτική επαλήθευση των αποτελεσμάτων. Παρ' όλο που δεν γνωρίζουμε το περιεχόμενο του σήματος, στην παραπάνω περίπτωση είναι εύκολο να καταλάβουμε ότι υπάρχει κάποιο λάθος στους ομιλητές #2 και #3 (το ίδιο ισχύει και για τα αποτελέσματα του Σχήματος 3.13). Έτσι είναι λογικό να υποθέσουμε ότι ο ίδιος ομιλητής έχει χωριστεί σε δύο ομάδες και να επαναλάβουμε τη διαδικασία θεωρώντας τρεις ομιλητές. Μία άλλη λύση είναι η χρήση μιας μεθόδου ομαδοποίησης που εκτιμά τον αριθμό των ομάδων (π.χ. [Pelleg and Moore, 2000]) και μία τρίτη είναι η εκτέλεση πολλών επαναλήψεων της ομαδοποίησης και αποδοχή της καλύτερης σύμφωνα με κάποιο κριτήριο επαλήθευσης (π.χ. πλάτος σιλουέτας). Στο συνοδευτικό λογισμικό έχουμε προτιμήσει την τελευταία πρόταση καθώς δεν απαιτείται παρέμβαση από μέρους μας όπως στην πρώτη και είναι σημαντικά πιο οικονομική υπολογιστικά από την δεύτερη.

3.2.3 diarx

Όπως αναφέραμε στον πρόλογο, στα πλαίσια αυτής της εργασίας αναπτύχθηκε το λογισμικό *diarx* [1] σε γλώσσα C++, με τη χρήση του οποίου έχουν γίνει πολλές δοκιμές για τη μέτρηση της αποδοτικότητας της μεθόδου.

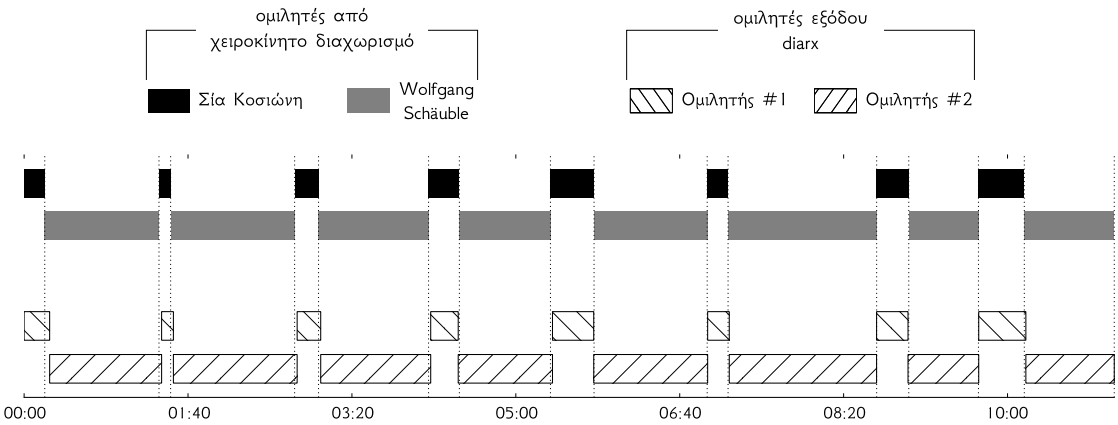
Η μέτρηση αυτή γίνεται με τη σύγκριση του διαχωρισμού που προτείνεται από το πρόγραμμα με έναν άλλον που έχει δημιουργηθεί χειροκίνητα και θεωρείται "πραγματικός". Τα εισαγωγικά μπαίνουν για να τονισθεί το γεγονός ότι ο χειροκίνητος διαχωρισμός είναι υποκειμενικός. Για παράδειγμα, μικρές παύσεις ανάμεσα σε προτάσεις μπορεί να θεωρηθούν από κάποιον ως περιοχές χωρίς ομιλία, ενώ από κάποιον άλλον ως περιοχές που "ανήκουν" σε κάποιον ομιλητή. Αντίστοιχες περιπτώσεις αποτελούν και τα τμήματα όπου ένας ομιλητής μιλάει μαζί με κάποιον άλλον για ένα μικρό χρονικό διάστημα κτλ. Γενικά τέτοιες καταστάσεις δεν αποτελούν πρόβλημα αφού συνήθως αφορούν πολύ μικρό ποσοστό ενός σήματος ομιλίας, αλλά σε ένα πλήρες σύστημα επεξεργασίας ομιλίας θα λαμβανόταν η απαιτούμενη μέριμνα για τέτοιες περιπτώσεις (π.χ. απόρριψη πλαισίων ομιλίας όπου δεν υπάρχει φωνητική δραστηριότητα).

Σημείωση 6. *Το λογισμικό diarx δεν αποτελεί σε καμία περίπτωση ένα τέτοιο πλήρες σύστημα, αλλά παρατηρήθηκε ότι η χρήση μεγάλων παραθύρων ανάλυσης και των μέσων όρων των χαρακτηριστικών έχει σαν συνέπεια την ακόμα μικρότερη επίδραση αυτών των περιπτώσεων στην απόδοσή του.*

Οι πρώτες δοκιμές του λογισμικού έγιναν σε δελτία ειδήσεων και συνεντεύξεις που μπορούν να βρεθούν εύκολα στο διαδίκτυο μέσα από τα αρχεία των τηλεοπτικών σταθμών. Αυτές οι κατηγορίες σημάτων είναι ιδανικές για εφαρμογή αυτόματου διαχωρισμού ομιλητών επειδή *έχει νόημα* να γίνει κάτι τέτοιο και ταυτόχρονα η δομή τους είναι συνήθως τέτοια που να μην παραβιάζονται σε μεγάλο βαθμό οι υποθέσεις που έχουμε κάνει. Ένα παράδειγμα είναι η συνέντευξη του Σχήματος 3.15. Ακόμα και με γυμνό μάτι μπορούμε να καταλάβουμε ότι ο αυτόματος διαχωρισμός είναι σωστός.

Η διαδικασία που ακολουθείται για τη μέτρηση της απόδοσης της μεθόδου έχει ως εξής:

[1] Τμήματα αυτού του λογισμικού μπορούν να βρεθούν ελεύθερα στο αποθετήριο git https://gntikos@bitbucket.org/gntikos/diarx.git



Σχήμα 3.15: Διαχωρισμός ομιλητών σε μία συνέντευξη

1. Σχηματίζεται ο πίνακας σύγχυσης των δύο ομαδοποιήσεων. Κάθε γραμμή αντιστοιχεί σε έναν ομιλητή από τη χειροκίνητη ομαδοποίηση και κάθε στήλη σε έναν ομιλητή από την αυτόματη ομαδοποίηση του λογισμικού.

2. Στο κάθε κελί (i, j) γράφεται η διάρκεια των τμημάτων στα οποία μιλάει ο ομιλητής i (από τη χειροκίνητη ομαδοποίηση) και ο ομιλητής j (από την αυτόματη).

3. Γίνεται αντιστοίχηση των δύο συνόλων ομιλητών με την "ουγγρική μέθοδο" [Bourgeois and Lassalle, 1971]. Για παράδειγμα, στην περίπτωση του Σχήματος 3.15 ο "Ομιλητής #1" αντιστοιχεί στη δημοσιογράφο και ο "Ομιλητής #2" στον πολιτικό που δίνει τη συνέντευξη.

4. Η απόδοση της ομαδοποίησης ποσοτικοποιείται με χρήση μέτρων όπως τα παρακάτω:

   - Ποσοστό Ακρίβειας Διαχωρισμού (Diarization Accuracy Rate - DAR) Αυτό είναι το βασικό μέτρο που χρησιμοποιούμε. Ορίζεται ως ο λόγος των τμημάτων που έχουν ομαδοποιηθεί *σωστά* προς το σύνολο των τμημάτων.
   - Κανονικοποιημένη Αμοιβαία Πληροφορία (Normalized Mutual Information - NMI_{max}) Ένα μέτρο πληροφορίας που συγκρίνει δύο κατατμήσεις του ίδιου συνόλου δεδομένων. Είναι φραγμένο στο διάστημα $[0, 1]$ και ισούται με 1 όταν οι δύο κατατμήσεις είναι ίδιες και με 0 όταν δεν περιέχουν

κοινή πληροφορία. Στο [Vinh et al., 2010] παρουσιάζεται μαζί με διάφορες παραλλαγές του.

- Μέση Καθαρότητα Ομάδας (Average Cluster Purity - ACP) Αυτό το μέτρο εκφράζει το κατά πόσο τα τμήματα που έχουν ταξινομηθεί σε μία ομάδα ανήκουν σε έναν ομιλητή. Με άλλα λόγια, η "καθαρότητα" μιας ομάδας έχει να κάνει με το αν περιέχει τμήματα από ομιλητές εκτός αυτού στον οποίον αντιστοιχεί. Συμβολίζοντας με n_{ij} το στοιχείο του πίνακα σύγχυσης στη θέση (i, j), με N το άθροισμα των στοιχείων του και με N_c, N_s τον αριθμό των ομάδων που δημιουργήθηκαν και τον πραγματικό αριθμό των ομιλητών αντίστοιχα, το μέτρο δίνεται από τη σχέση

$$ACP = \frac{1}{N} \sum_{j=1}^{N_c} \max_{i=1 \ldots N_s} n_{ij}$$

- Μέση Καθαρότητα Ομιλητή (Average Speaker Purity - ASP) Με εντελώς αντίστοιχο τρόπο όπως στο προηγούμενο μέτρο, η μέση καθαρότητα ομιλητή εκφράζει το κατά πόσο ένας ομιλητής έχει αντιστοιχηθεί σε μία μόνο ομάδα. Ορίζεται ως

$$ASP = \frac{1}{N} \sum_{j=1}^{N_s} \max_{j=1 \ldots N_c} n_{ij}$$

Στο παράδειγμα του Σχήματος 3.15 οι τιμές που παίρνουν τα διάφορα μέτρα είναι

αρ. ομιλ.	διάρκεια	$DAR(\%)$	$NMI_{max}(\%)$	$ASP(\%)$	$ACP(\%)$
2	11:05	97.94%	81.06%	97.94%	97.94%

Επίσης, το πρόγραμμα δοκιμάστηκε σε ένα σύνολο 53 πολιτικών εκπομπών του γαλλικού Canal 9 [Vinciarelli et al., 2009]. Οι εκπομπές αυτές έχουν τη μορφή debate όπου οι ομιλητές επιχειρηματολογούν σε θέματα που θέτει ο παρουσιαστής και έχουν μια μέση διάρκεια 35 περίπου λεπτών. Το ίδιο σύνολο εκπομπών χρησιμοποιήθηκε για την αξιολόγηση της μεθόδου στο [Giannakopoulos and Petridis, 2012], οπότε η επιλογή του εδώ χρησιμεύει και σαν αξιολόγηση της υλοποίησής μας. Όσον αφορά τις δοκιμές, κάποιες παρατηρήσεις είναι απαραίτητες:

- Τα αρχεία περιέχουν το μουσικό θέμα της εκπομπής στην αρχή και το τέλος τους, κάτι που είναι σύνηθες ιδιαίτερα για τηλεοπτικές εκπομπές. Αυτά τα τμήματα αφαιρέθηκαν ώστε να ικανοποιείται η προϋπόθεση ότι το σήμα περιέχει *μόνο ομιλία*.

- Ο υπόχωρος υπολογίστηκε με τη μέθοδο FLsD από το υπόλοιπο σήμα. Παρ' όλα αυτά, από τη στιγμή που δεν υπάρχει μέριμνα για περιπτώσεις όπου δύο ομιλητές μιλούν ταυτόχρονα, η αξιολόγηση έγινε αγνοώντας αυτά τα τμήματα. Παρόμοια τακτική ακολουθείται και στο [Giannakopoulos and Petridis, 2012] αν και εκεί αγνοούνται και κατά τον υπολογισμό του υπόχωρου.

- Η πιο σοβαρή έλλειψη του λογισμικού είναι ότι δεν υπολογίζει αυτόματα τον αριθμό των ομιλητών. Παρ' όλο που όπως περιγράψαμε υπάρχουν τρόποι για να γίνει αυτό, δεν μπορέσαμε να καταλήξουμε σε κάποια ικανοποιητική υλοποίηση. Έτσι, ο αριθμός των ομιλητών θεωρήθηκε γνωστός σε όλες τις περιπτώσεις.

- Όλα τα παραπάνω γίνονται με αυτόματο τρόπο, με scripts που μπορούν να βρεθούν στο φάκελο utils του προγράμματος.

Τα αποτελέσματα των δοκιμών παρουσιάζονται στον πίνακα 3.1:

Α/Α	αρχείο	αρ. ομιλ.	διάρκεια	$DAR(\%)$	$NMI_{max}(\%)$	$ASP(\%)$	$ACP(\%)$
1	05-09-28	3	0:40:11	78.20	55.29	78.21	86.82
2	05-10-05	3	0:34:50	95.59	80.12	95.60	95.59
3	05-10-12	3	0:30:48	88.48	69.83	88.48	88.48
4	05-10-19	3	0:27:58	96.38	82.00	96.39	96.38
5	05-10-26	3	0:29:27	79.47	58.68	79.47	83.92
6	05-11-02	3	0:28:28	95.24	79.67	95.24	95.24
7	05-11-09	3	0:31:23	85.89	58.37	85.89	85.89
8	05-11-16	3	0:28:43	70.49	48.00	70.49	83.77
9	05-11-23	3	0:30:11	70.15	52.83	70.79	85.08
10	05-12-07	3	0:30:35	96.95	85.75	96.96	96.95
11	05-12-14	4	0:32:02	76.65	66.98	83.91	88.75

12	06-01-11	3	0:29:14	95.21	79.10	95.22	95.21
13	06-01-18	3	0:30:21	86.74	62.25	86.74	86.74
14	06-02-08	4	0:28:15	94.89	83.62	94.89	94.89
15	06-02-15	3	0:31:06	81.18	56.70	81.44	81.34
16	06-02-22	3	0:30:12	96.67	84.94	96.67	96.67
17	06-03-22	3	0:30:17	87.25	61.58	87.26	87.25
18	06-03-29	3	0:29:56	89.81	69.30	89.81	89.81
19	06-04-05	5	0:44:55	91.71	77.28	91.71	91.71
20	06-04-12	5	0:41:52	85.14	70.55	89.24	85.14
21	06-04-19	4	0:38:54	80.36	59.42	80.37	80.36
22	06-05-03	5	0:34:05	95.11	83.78	95.11	95.11
23	06-05-10	5	0:40:48	32.23	0.09	99.94	32.23
24	06-05-17	5	0:43:59	86.78	72.20	88.69	86.78
25	06-05-24	5	0:36:20	89.00	75.62	89.01	89.00
26	06-05-31	5	0:40:42	87.54	70.61	87.54	87.54
27	06-06-07	5	0:45:29	94.84	84.91	94.84	94.84
28	06-06-14	4	0:41:45	94.71	80.98	94.72	94.71
29	06-09-06	5	0:37:11	85.09	65.81	86.74	85.09
30	06-09-13	5	0:39:33	90.52	74.52	90.52	90.52
31	06-09-20	5	0:37:29	90.12	75.73	90.12	90.12
32	06-09-27	5	0:34:35	81.95	59.93	81.96	81.95
33	06-10-04	5	0:37:32	91.92	80.91	91.93	91.92
34	06-10-11	5	0:36:04	91.48	78.20	91.49	91.48
35	06-10-18	5	0:38:13	91.31	75.70	91.31	91.31
36	06-11-08	5	0:36:14	90.79	76.15	90.79	90.79
37	06-11-15	5	0:39:01	89.69	77.63	89.69	89.69
38	06-11-22	5	0:37:35	93.52	82.29	93.53	93.52
39	06-11-29	5	0:40:07	92.82	80.85	92.82	92.82
40	06-12-06	5	0:36:34	90.80	74.94	90.80	90.80
41	06-12-13	5	0:36:50	93.66	80.81	93.66	93.66
42	06-12-20	5	0:36:42	83.98	65.12	88.63	83.98
43	07-01-17	5	0:36:08	91.18	75.45	91.18	91.18
44	07-01-31	5	0:36:54	88.51	73.87	88.51	88.51

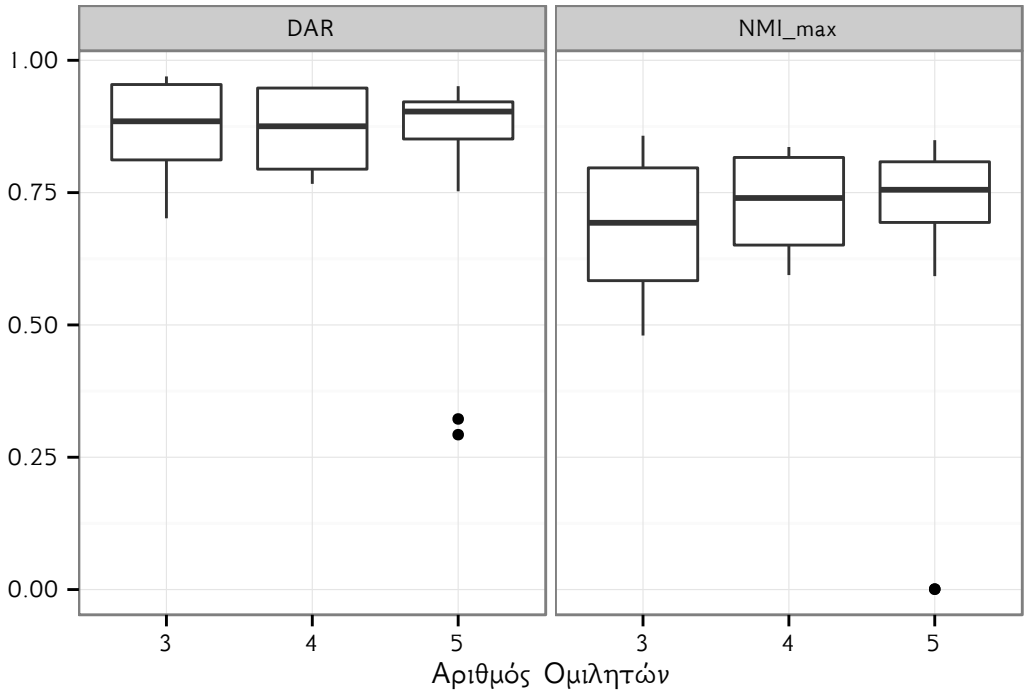
45	07-02-14	5	0:37:39	94.12	82.98	94.12	94.12
46	07-02-21	5	0:37:11	75.70	59.22	85.22	75.70
47	07-02-28	5	0:37:22	94.05	82.54	94.05	94.05
48	07-03-21	3	0:29:03	95.42	79.34	95.42	95.42
49	07-03-28	5	0:38:49	88.24	72.57	88.33	88.24
50	08-01-09	5	0:34:29	93.49	81.84	93.77	93.49
51	08-01-15	5	0:37:35	29.26	0.07	99.94	29.26
52	08-01-23	5	0:40:40	80.14	61.92	91.75	80.19
53	08-01-30	5	0:37:18	75.25	59.59	89.00	75.25
Μέση Τιμή				86.15	69.69	89.81	87.15
Διασπορά				1.69	2.88	0.38	1.5

Πίνακας 3.1: Αποτελέσματα δοκιμών του diarx σε αρχεία του σταθμού Canal 9

Σχεδόν σε κάθε περίπτωση τα αποτελέσματα είναι πολύ ικανοποιητικά και μάλιστα στο Σχήμα 3.16 βλέπουμε ότι αυτό συμβαίνει ανεξάρτητα από τον αριθμό των ομιλητών.

Τα αποτελέσματα που προέκυψαν είναι σύμφωνα με αυτά του [Giannakopoulos and Petridis, 2012] και ο χρόνος εκτέλεσης των δοκιμών (04'59" για το διαχωρισμό και την αξιολόγηση αρχείων συνολικής διάρκειας 31ω29'34") είναι ικανοποιητικός. Συνεπώς, η υλοποίηση του diarx μπορεί να θεωρηθεί επιτυχής και θα μπορούσε να χρησιμοποιηθεί και σε άλλα πραγματικά δεδομένα.

Τελειώνοντας, μπορούμε να αναφερθούμε στο επόμενο βήμα της ανάπτυξης της μεθόδου, που είναι η επέκτασή της ώστε να μπορεί να χρησιμοποιηθεί σε εφαρμογές πραγματικού χρόνου (π.χ. διαχωρισμό σε πολυμεσικές ροές). Αξίζει να σημειωθεί ότι επιχειρώντας αυτή την επέκταση, επανέρχεται το πρόβλημα εκτίμησης του αριθμού ομιλητών: στην περίπτωση που παρουσιάσαμε, η λάθος εκτίμηση οδηγεί σε περιπτώσεις που το λάθος είτε μπορεί εύκολα να εντοπισθεί (όπως π.χ. στο Σχ. 3.14) είτε επηρεάζει ένα περιορισμένο τμήμα του σήματος. Σε μία εφαρμογή όπου ο αριθμός των ομιλητών πρέπει να εκτιμάται συνεχώς, η αδυναμία συνεπούς και αξιόπιστης εκτίμησης είναι αναμενόμενο να έχει πολύ μεγαλύτερες συνέπειες. Πιθανές τροποποιήσεις προς αυτή την κατεύθυνση είναι η ένταξη χρονικών περιορισμών κατά τη διαδικασία ομαδοποίησης ώστε να αντανακλάται (όπως και κατά το σχηματισμό νημάτων κλάσης)



Σχήμα 3.16

η ροή ενός διαλόγου, ή ακόμα και η χρήση μιας μη-παραμετρικής μεθόδου ανίχνευσης σημείων αλλαγής. Σε κάθε περίπτωση, πιστεύουμε πως η συμβολή της μεθόδου είναι σημαντική, αφού αντιμετωπίζει με αποτελεσματικό και θεωρητικά έγκυρο τρόπο μια βασική αδυναμία των περισσότερων μεθόδων και εκτιμούμε πως η χρήση της θα βοηθήσει στην ανάπτυξη αποτελεσματικότερων συστημάτων διαχωρισμού ομιλητών.

Βιβλιογραφία

[Adami et al., 2002] Adami, A. G., Kajarekar, S. S., and Hermansky, H. (2002). A new speaker change detection method for two-speaker segmentation. In *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, volume 4, pages IV–3908–IV–3911.

[Ajmera et al., 2004] Ajmera, J., McCowan, I., and Bourlard, H. (2004). Robust speaker change detection. *Signal Processing Letters, IEEE*, 11(8):649–651.

[Ajmera and Wooters, 2003] Ajmera, J. and Wooters, C. (2003). A Robust Speaker Clustering Algorithm. In *IEEE Automatic Speech Recognition Understanding Workshop*.

[Anderson, 1984] Anderson, J. C. (1984). *Speech analysis/synthesis based on perception*. PhD thesis, Massachusetts Institute of Technology. Dept. of Electrical Engineering and Computer Science.

[Arthur and Vassilvitskii, 2007] Arthur, D. and Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

[Bachu et al., 2010] Bachu, R., Kopparthi, S., Adapa, B., and Barkana, B. (2010). Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy. In *Advanced Techniques in Computing Sciences and Software Engineering*, pages 279–282. Springer Netherlands.

[Bahmani et al., 2012] Bahmani, B., Moseley, B., Vattani, A., Kumar, R., and Vassilvitskii, S. (2012). Scalable k-means++. *Proc. VLDB Endow.*, 5(7):622–633.

[Belhumeur et al., 1997] Belhumeur, P., Hespanha, J., and Kriegman, D. (1997). Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720.

[Bellman, 1961] Bellman, R. E. (1961). *Adaptive control processes - A guided tour*. Princeton University Press.

[Bezdek et al., 1984] Bezdek, J., Ehrlich, R., and Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3):191–203.

[Bezdek, 1981] Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA.

[Bogert et al., 1963] Bogert, B., Healy, M., and Tukey, J. (1963). The quefrency alanysis of time series for echoes: Cepstrum, Pseudo-Autocovariance, Cross-Cepstrum and Saphe Cracking. In *Proc. Symp. on Time Series Analysis*, pages 209–243.

[Bourgeois and Lassalle, 1971] Bourgeois, F. and Lassalle, J.-C. (1971). An extension of the munkres algorithm for the assignment problem to rectangular matrices. *Commun. ACM*, 14(12):802–804.

[Brockwell and Davis, 1986] Brockwell, P. J. and Davis, R. A. (1986). *Time series: theory and methods*. Springer-Verlag New York, Inc., New York, NY, USA.

[Castaldo et al., 2008] Castaldo, F., Colibro, D., Dalmasso, E., Laface, P., and Vair, C. (2008). Stream-based speaker segmentation using speaker factors and eigenvoices. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4133–4136.

[Chen, 2000] Chen, L. (2000). A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, 33(10):1713–1726.

[Chen and Gopalakrishnan, 1998] Chen, S. S. and Gopalakrishnan, P. S. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 127–132.

[Cheng et al., 2010] Cheng, S.-S., min Wang, H., and Fu, H.-C. (2010). Bic-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(1):141–157.

[Chu et al., 2009] Chu, S., Tang, H., and Huang, T. (2009). Fishervoice and semi-supervised speaker clustering. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4089–4092.

[Dagher, 2010] Dagher, I. (2010). Incremental pca-lda algorithm. In *Computational Intelligence for Measurement Systems and Applications (CIMSA), 2010 IEEE International Conference on*, pages 97–101.

[Davis and Mermelstein, 1980] Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366.

[Durbin, 1960] Durbin, J. (1960). The Fitting of Time-Series Models. *Review of the International Statistical Institute*, 28(3).

[Elkan, 2003] Elkan, C. (2003). Using the triangle inequality to accelerate k-means.

[Estivill-Castro, 2002] Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. *SIGKDD Explor. Newsl.*, 4(1):65–75.

[Fant, 1960] Fant, G. (1960). *Acoustic Theory of Speech Production*. Mouton.

[Flanagan, 1972] Flanagan, J. (1972). *Speech analysis synthesis and perception*. Kommunikation und Kybernetik in Einzeldarstellungen. Springer-Verlag.

[Foley and Sammon, 1975] Foley, D. and Sammon, J. (1975). An optimal set of discriminant vectors. *Computers, IEEE Transactions on*, C-24(3):281–289.

[Giannakopoulos and Petridis, 2012] Giannakopoulos, T. and Petridis, S. (2012). Fisher linear semi-discriminant analysis for speaker diarization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(7):1913–1922.

[Guha et al., 2003] Guha, S., Meyerson, A., Mishra, N., and Motwani, R. (2003). Clustering data streams: Theory and practice. *IEEE TKDE*, 15:515–528.

[Hamerly, 2010] Hamerly, G. (2010). Making k-means even faster. In *Proceedings of the 2010 SIAM international conference on data mining (SDM 2010)*.

[Hotelling, 1933] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, 24.

[Itakura, 1975] Itakura, F. (1975). Line spectrum representation of linear predictor coefficients of speech signals. 57(S1).

[Kabal and Ramachandran, 1986] Kabal, P. and Ramachandran, R. (1986). The computation of line spectral frequencies using chebyshev polynomials. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(6):1419–1426.

[Kotti et al., 2008] Kotti, M., Benetos, E., and Kotropoulos, C. (2008). Computationally efficient and robust bic-based speaker segmentation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(5):920–933.

[Kotti et al., 2006] Kotti, M., Martins, L. G. P. M., Benetos, E., Cardoso, J. S., and Kotropoulos, C. (2006). Automatic speaker segmentation using multiple features and distance measures: A comparison of three approaches. In *in IEEE 2006 International Conference on Multimedia & Expo (ICME2006)*.

[Levinson, 1947] Levinson, N. (1947). The Wiener RMS (root mean square) error criterion in filter design and prediction. *J. Math. Phys.*, 25(4):261–278.

[Li, 1996] Li, R.-C. (1996). Relative perturbation theory: (ii) eigenspace and singular subspace variations. *SIAM J. Matrix Anal. Appl*, 20:471–492.

[Lloyd, 1982] Lloyd, S. (1982). Least squares quantization in PCM. *Information Theory, IEEE Transactions on*, 28(2):129–137.

[Milner, 2009] Milner, G. (2009). *Perfecting Sound Forever: An Aural History of Recorded Music*. Faber & Faber.

[O'Callaghan et al., 2001] O'Callaghan, L., Mishra, N., Meyerson, A., Guha, S., and Motwani, R. (2001). Streaming-data algorithms for high-quality clustering. In *Proceedings of IEEE International Conference on Data Engineering*, page 685.

[Oppenheim and Schafer, 2004] Oppenheim, A. and Schafer, R. (2004). From frequency to quefrency: a history of the cepstrum. *Signal Processing Magazine, IEEE*, 21(5):95–106.

[Oppenheim, 1965] Oppenheim, A. V. (1965). Superposition in a class of nonlinear systems. Technical Report 432, Massachusetts Institute of Technology.

[Pearson, 1901] Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572.

[Pelleg and Moore, 2000] Pelleg, D. and Moore, A. (2000). X-means: Extending k-means with efficient estimation of the number of clusters. In *In Proceedings of the 17th International Conf. on Machine Learning*, pages 727–734. Morgan Kaufmann.

[Rabiner and Schafer, 1978] Rabiner, L. R. and Schafer, R. W. (1978). *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ.

[Schafer, 1968] Schafer, R. (1968). *Echo Removal by Discrete Generalized Linear Filtering*. Massachusetts Institute of Technology. Research Laboratory of Electronics. Technical report. Massachusetts Institute of Technology, Research Laboratory of Electronics.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

[Shannon, 1949] Shannon, C. (1949). Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21.

[Soong and Juang, 1984] Soong, F. and Juang, B.-H. (1984). Line spectrum pair (lsp) and speech data compression. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.*, volume 9, pages 37–40.

[Stevens et al., 1937] Stevens, S. S., Je, and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude of pitch. *J. Acoust Soc Amer*, 8:185–190.

[Tang et al., 2005] Tang, Y., Sun, F., and Sun, Z. (2005). Improved validation index for fuzzy clustering. In *American Control Conference, 2005. Proceedings of the 2005*, pages 1120–1125 vol. 2.

[Tritschler and Gopinath, 1999] Tritschler, A. and Gopinath, R. A. (1999). Improved speaker segmentation and segments clustering using the bayesian information criterion. In *EUROSPEECH'99*.

[Tzanetakis and Cook, 2002] Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *Speech and Audio Processing, IEEE Transactions on*, 10(5):293–302.

[Umesh et al., 1999] Umesh, S., Cohen, L., and Nelson, D. (1999). Fitting the mel scale. In *Proceedings of the Acoustics, Speech, and Signal Processing, 1999. on 1999 IEEE International Conference - Volume 01*, ICASSP '99, pages 217–220. IEEE Computer Society.

[van der Heijden et al., 2004] van der Heijden, F., Duin, R., de Ridder, D., and Tax, D. M. J. (2004). *Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB*. Wiley, 1 edition.

[Vendramin et al., 2010] Vendramin, L., Campello, R. J. G. B., and Hruschka, E. R. (2010). Relative clustering validity criteria: A comparative overview. *Stat. Anal. Data Min.*, 3(4):209–235.

[Vinciarelli et al., 2009] Vinciarelli, A., Dielmann, A., Favre, S., and Salamin, H. (2009). Canal9: A database of political debates for analysis of social interactions. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–4.

[Vinh et al., 2010] Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.*, 11:2837–2854.

[Zadeh, 1965] Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8(3):338 – 353.

[Zhang et al., 2013] Zhang, L.-H., Liao, L.-Z., and Ng, M. (2013). Superlinear convergence of a general algorithm for the generalized foley–sammon discriminant analysis. *Journal of Optimization Theory and Applications*, 157(3):853–865.

[Zhou and Hansen, 2005] Zhou, B. and Hansen, J. (2005). Efficient audio stream segmentation via the combined t^2 statistic and bayesian information criterion. *Speech and Audio Processing, IEEE Transactions on*, 13(4):467–474.