Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Τομεας Σηματων, Ελεγχου και Ρομποτικης

Αναγνώριση συναισθήματος μέσω φωνής

Διπλωματικη Εργασια

του

Γιαννούλη Κ. Παναγιώτη

Επιβλέπων: Ποταμιάνος Γεράσιμος
Αν. Καθηγητής Πανεπιστήμιο Θεσσαλίας

Αθήνα, Σεπτέμβριος 2012

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Αναγνώριση συναισθήματος μέσω φωνής

Διπλωματικη Εργασια

του

Γιαννούλη Κ. Παναγιώτη

Επιβλέπων: Ποταμιάνος Γεράσιμος
Αν. Καθηγητής Πανεπιστήμιο Θεσσαλίας

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 20η Σεπτεμβρίου 2012.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Ποταμιάνος Γεράσιμος	Μαραγκός Πέτρος	Τζαφέστας Κώστας
Αν. Καθηγητής Π.Θ.	Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2012

(Υπογραφή)

.....

Γιαννουλης Παναγιωτης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Ευχαριστίες

Θα ήθελα καταρχήν να ευχαριστήσω θερμά τον Αν. καθηγητή του Π.Θ. και συνεργαζόμενο ερευνητή στον Δημόκριτο κ. Γεράσιμο Ποταμιάνο, για την επίβλεψη αυτής της διπλωματικής εργασίας. Η καθοδήγηση του ήταν συνεχής και ιδιαίτερα πολύτιμη, καθόλη την διάρκεια της ερευνητικής μας προσπάθειας.

Θα ήθελα επίσης να ευχαριστήσω τον καθηγητή του Ε.Μ.Π. κ. Πέτρο Μαραγκό, καθώς ήταν η βασική αιτία για να ξεκινήσω να ασχολούμαι με την Ψηφιακή Επεξεργασία Σημάτων και με την Αναγνώριση Προτύπων, και ως συνέπεια να επιλέξω ως θέμα της διπλωματικής μου την Αναγνώριση συναισθήματος μέσω φωνής.

Τέλος, να ευχαριστήσω τον Δημήτρη Δημητριάδη, για την παραχώρηση του Matlab κώδικα σχετικά με τα χαρακτηριστικά AM-FM.

Λέξεις Κλειδιά

Emotion recognition, Glottal flow, AM-FM features, Feature selection, GMM, SVM

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Η αναγνώριση συναισθήματος αποτελεί ενεργό κομμάτι της έρευνας τα τελευταία χρόνια έχοντας ως κύριο στόχο την βελτίωση της επικοινωνίας μεταξύ ανθρώπου και μηχανής. Το συναίσθημα παίζει βασικό ρόλο στις διαπροσωπικές σχέσεις των ανθρώπων και περιέχει σημαντική πληροφορία για την κατάσταση του ατόμου, τις προθέσεις του, και τις πράξεις του στο άμεσο μέλλον. Για να γίνει πιο άνετη, χρήσιμη, και ευχάριστη η επικοινωνία μεταξύ ανθρώπου-υπολογιστή, θα πρέπει ο υπολογιστής να αντιλαμβάνεται τα συναισθήματα του ανθρώπου και να αντιδρά κατάλληλα. Αρκετοί μάλιστα υποστηρίζουν ότι η συναισθηματική ευφυία των υπολογιστών είναι εξίσου σημαντική με την υπολογιστική, αφού είναι απαραίτητη για να διαπιστωθούν οι προτιμήσεις του κάθε ανθρώπου και να υπάρξει προσαρμογή που θα δώσει πραγματική διαδραστικότητα στην επικοινωνία. Παρακάτω αναφέρουμε μερικές εφαρμογές της αναγνώρισης συναισθήματος μέσω φωνής, όπως έχουν αναφερθεί σε υπάρχουσες έρευνες:

- Ρομποτ για χρήση στο σπίτι αποκτούν συναισθηματική νοημοσύνη και γίνονται πιο φιλικά στους ανθρώπους [1] .

- Είναι γνωστό ότι πολλά αυτοκινητιστικά ατυχήματα έχουν συμβεί σαν συνέπεια της έντονης ψυχολογικής φόρτισης του οδηγού. Για την αποτροπή τέτοιων κινδύνων, ένα σύστημα αναγνώρισης συναισθήματος μέσω φωνής, θα μπορούσε να εντοπίζει περιπτώσεις θυμού, άγχους και απογοήτευσης και με κατάλληλα μηνύματα να προειδοποιεί και να ηρεμεί τον οδηγό για την δική του ασφάλεια [2, 3].

- Στον τομέα των τηλεπικοινωνιών έχει αναπτυχθεί εφαρμογή με το όνομα *Voice Driven Emotion Recognizer Mobile-phone (VDERM)* [4], η οποία στοχεύει στην μετάδοση μη λεκτικής πληροφορίας μέσω τηλεφώνων, με περισσότερο αποδοτικό και οικονομικό τρόπο σε σχέση με την μετάδοση βίντεο. Συγκεκριμένα αντί να μεταφέρεται ολόκληρη η εικόνα του ομιλητή, γίνεται αναγνώριση του συναισθήματος του μέσω φωνής, και μεταβάλλεται ανάλογα η έκφραση ενός προσώπου στο τηλέφωνο του δέκτη. Υπάρχει έτσι μεγάλο κέρδος στον χρόνο και το κόστος μεταφοράς του σήματος.

- Στον τομέα της ιατρικής και της ψυχολογίας, η μελέτη της επίδρασης ψυχολογικών

ασταθειών, όπως είναι η κατάθλιψη, στην φωνή θα μπορούσε να ενισχύσει σημαντικά την διάγνωση των γιατρών [5, 6].

Στις επόμενες ενότητες παρουσιάζουμε την ανάλυση των συναισθημάτων από την πλευρά της επιστήμης της ψυχολογίας καθώς και τον τρόπο με τον οποίο μπορεί να πραγματοποιηθεί η αναγνώριση συναισθήματος μέσω φωνής.

1.1 Το συναίσθημα στην ψυχολογία

Στην ανθρώπινη επικοινωνία μπορεί κανείς να εντοπίσει δύο βασικούς διαύλους. Στον πρώτο μεταδίδονται σαφή μηνύματα (όπως λέξεις, προτάσεις, χειρονομίες κλπ.), ενώ στο δεύτερο υποννοούμενα μηνύματα που αντικατοπτρίζουν τις σκέψεις, προθέσεις, και την ψυχολογική κατάσταση των ανθρώπων κάθε στιγμή. Για την αποσαφήνιση του δεύτερου καναλιού, το οποίο περιλαμβάνει την κατανόηση των συναισθημάτων, συνεχίζονται πολλές μελέτες.

Τα συναισθήματα αναφέρονται στην ψυχική κατάσταση που βιώνει το άτομο. Ο κύκλος τους αποτελείται από μια αλληλουχία γεγονότων που ξεκινάει με το ερέθισμα το οποίο δημιουργεί ψυχολογικές μεταβολές, συνεχίζει με την αντίδραση σε αυτό το ερέθισμα, και τελειώνει με μια συγκεκριμένη ενέργεια που εξυπηρετεί κάποιο στόχο. Τα συναισθήματα λοιπόν δεν συμβαίνουν ποτέ μόνα τους, αλλά αποτελούν την αντίδραση σε καταστάσεις της ζωής και με την σειρά τους, όταν εξωτερικεύονται, προκαλούν κι αυτά νέες αντιδράσεις.

Η σημασία των συναισθημάτων είναι θεμελιώδης στην ζωή των ανθρώπων καθώς ρυθμίζουν την ψυχική τους υγεία και παίζουν κεντρικό ρόλο στις διαπροσωπικές τους σχέσεις. Ακόμα τα συναισθήματα επηρεάζουν την λογική σκέψη, αφού πολλές φορές η συναισθηματική ευφυία είναι απαραίτητη για την λήψη λογικών αποφάσεων.

Στο χώρο της ψυχολογίας υπάρχουν δύο βασικοί τρόποι προσέγγισης των συναισθημάτων: Η πρώτη θεωρία υποστηρίζει την ύπαρξη διακριτών βασικών συναισθημάτων, ενώ η δεύτερη τα αναπαριστά ως σημεία πάνω σε διπολικούς άξονες 2 ή 3 διαστάσεων [7].

1.1.1 Βασικά συναισθήματα

Η θεωρία αυτή υποστηρίζει ότι υπάρχουν ορισμένα βασικά συναισθήματα που παρατηρούνται σε όλους τους πολιτισμούς παγκοσμίως. Τα συναισθήματα αυτά είναι διακριτά και το καθένα έχει μοναδική φυσιολογική διέγερση και τρόπο εκδήλωσης. Αν και υπάρχουν διαφωνίες για το ποιά είναι ακριβώς αυτά τα βασικά συναισθήματα, στις περισσότερες έρευνες θεωρούνται ως βασικά συναισθήματα η χαρά, τη θλίψη, ο θυμός και ο φόβος.

Για να ερμηνεύσουν την μεγάλη ποικιλία των διαφορετικών συναισθημάτων που υπάρχουν, οι υποστηρικτές αυτής της θεωρίας πιστεύουν πως τα βασικά συναισθήματα μπορούν να συγχωνευθούν έτσι ώστε να προκύψουν παράγωγα συναισθήματα. Η θεωρία αυτή είναι παρόμοια με την παραγωγή του φάσματος των χρωμάτων από τα 3 βασικά χρώματα. Στο σχήμα 1.1 φαίνεται ο τροχός συναισθημάτων που κατασκεύασε ο Plutchik [8] θέλωντας να δείξει τα βασικά συναισθήματα και τα παράγωγα που προκύπτουν από τις διάφορες αναμίξεις.

1.1.2 Διαστάσεις των συναισθημάτων

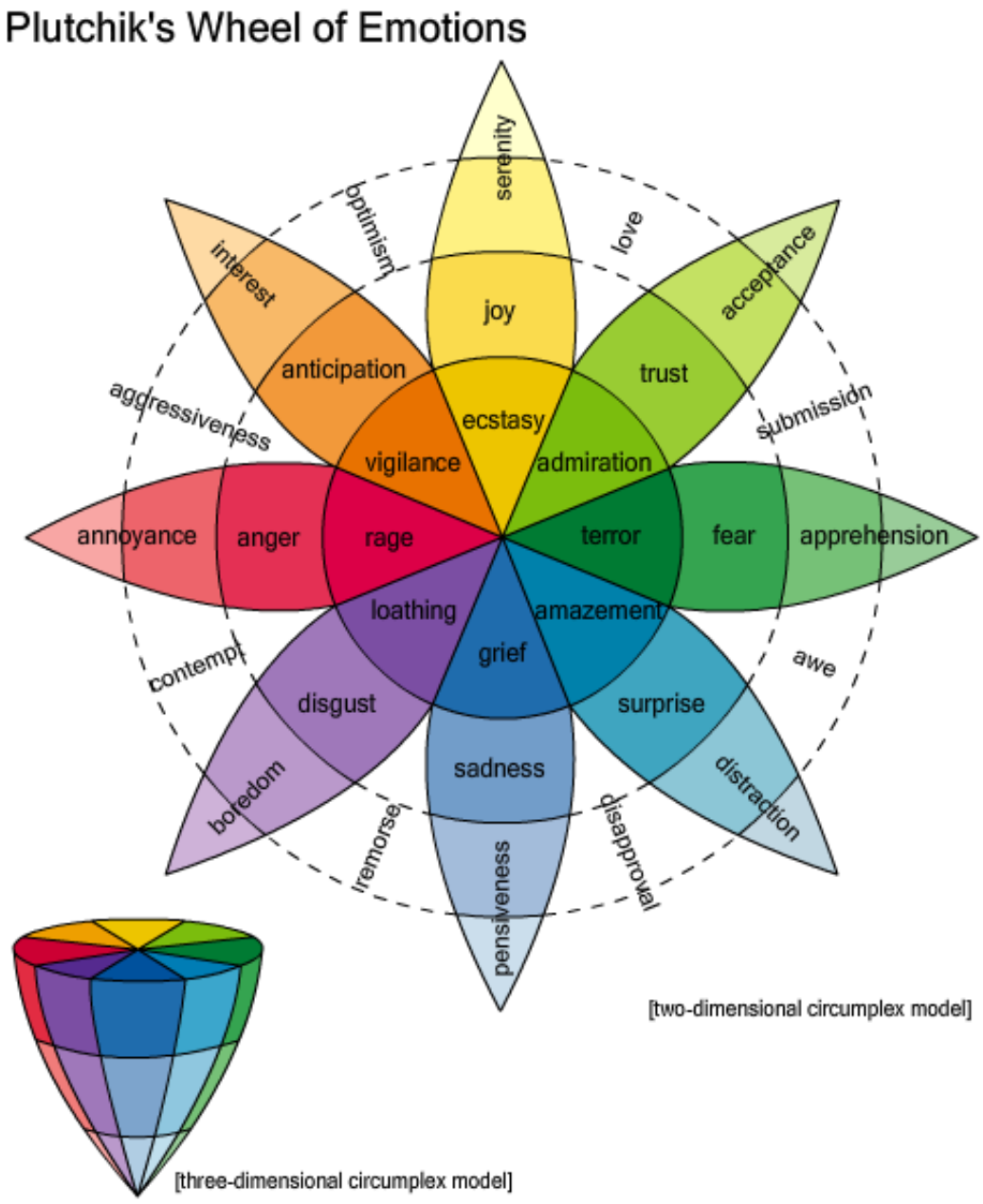
Ορισμένοι ερευνητές θέλησαν να εξαλείψουν τα μειονεκτήματα της προηγούμενης προσέγγισης, τα κυριότερα από τα οποία είναι η υποκειμενικότητα στην επιλογή των βασικών συναισθημάτων και η μη επαρκής εξήγηση για τον τρόπο ανάμειξης τους. Πρότειναν λοιπόν μια νέα παραμετρική προσέγγιση, ώστε να αναπαραστήσουν με πιο αξιόπιστο τρόπο το φάσμα των συναισθημάτων. Είναι γεγονός ότι τα βασικά συναισθήματα διαφέρουν ως προς την ένταση, το βαθμό ευχαρίστησης και τον βαθμό ενεργοποίησης. Είναι επομένως λογική προσέγγιση να θεωρηθούν τα συναισθήματα ως συναρτήσεις των τριών αυτών παραμέτρων και να αναπαρασταθούν ως σημεία σε τρεις διαστάσεις: ευχαρίστηση, διέγερση και δραστικότητα (valence, arousal, potency). Η ευχαρίστηση είναι η πιο σημαντική ίσως διάσταση, καθώς μας πληροφορεί για το αν το συναίσθημα είναι θετικό ή αρνητικό. Η διέγερση είναι ένα μέτρο της έντασης του συναισθήματος και της ψυχολογικής εμπλοκής του ατόμου σε αυτό, και τέλος, η δραστικότητα αντιπροσωπεύει την αίσθηση ελέγχου του ατόμου στο συναίσθημα. Τα δύο πρώτα χαρακτηριστικά είναι τα πιο σημαντικά και συχνά χρησιμοποιούμενα, γι΄ αυτό και υπάρχουν και αναπαραστάσεις σε δύο άξονες (Σχήμα 1.2).

1.2 Αναγνώριση συναισθημάτων μέσω φωνής

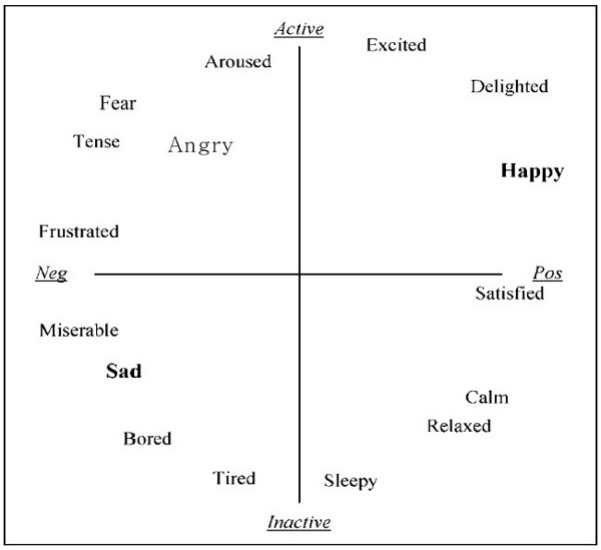
Δεδομένου ότι οι άνθρωποι μεταφέρουν με την φωνή πληροφορία που μαρτυρά την συναισθηματική τους κατάσταση, γεννιέται το ερώτημα πως θα μπορούσαμε να αναγνωρίσουμε με την βοήθεια ενός ηλεκτρονικού υπολογιστή συναισθήματα από στιγμιότυπα φωνής. Η απάντηση είναι αντιμετωπίζοντας το ζήτημα σαν ένα πρόβλημα αναγνώρισης προτύπων.

Ο κλάδος της αναγνώρισης προτύπων αναπτύχθηκε κυρίως την δεκαετία του 1960. Στόχος της είναι η εκπαίδευση ενός συστήματος με μεθόδους εκμάθησης μηχανής (machine learning) με βάση κάποια γνωστά αντικείμενα κάποιου είδους, έτσι ώστε να είναι δυνατή στη συνέχεια η αυτόματη και σωστή αναγνώριση νέων-άγνωστων αντικειμένων αυτού του είδους. Για την εκπαίδευση του συστήματος απαιτείται να υπάρχει μια βάση από δεδομένα, που στην περίπτωση μας είναι ένα σύνολο από ηχογραφημένα στιγμιότυπα φωνής με γνωστό συναισθηματικό περιεχόμενο. Τα βασικά βήματα της μεθοδολογίας αναγνώρισης προτύπων είναι τα παρακάτω:

- Εξαγωγή χαρακτηριστικών από τα δείγματα της βάσης.
- Επιλογή των καταλληλότερων χαρακτηριστικών.
- Επιλογή συστήματος-ταξινομητή, και εκπαίδευση του με βάση τα χαρακτηριστικά των δειγμάτων.
- Κατηγοριοποίηση νέων δειγμάτων από το σύστημα, αφού προηγηθεί η εξαγωγή αντίστοιχων χαρακτηριστικών από αυτά.



Σχήμα 1.1: Ο τροχός συναισθημάτων του Plutchik [8]



Σχήμα 1.2: Συναισθήματα στις δύο διαστάσεις [9]

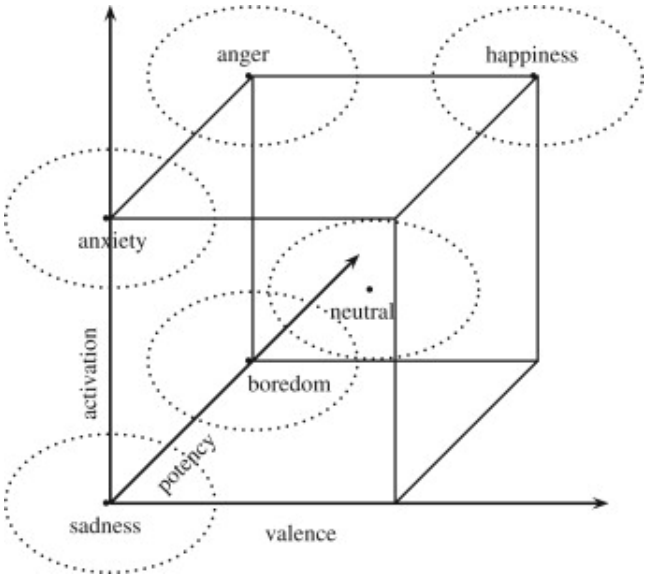
Στο σχήμα 1.4 φαίνεται πιο παραστατικά η ακολουθία των βασικών βημάτων.

Για τον τρόπο με τον οποίο καταλήγουν τα δείγματα φωνής των ατόμων αποθηκευμένα σε βάσεις δεδομένων στον υπολογιστή, μπορούμε να πούμε τα εξής. Αρχικά η φωνή που αποτελείται από κύματα πίεσης του αέρα, συλλαμβάνεται (ηχογραφείται) από το μικρόφωνο και μετατρέπεται σε ηλεκτρικό σήμα. Στη συνέχεια το ηλεκτρικό αυτό σήμα λαμβάνεται από τον υπολογιστή και υφίσταται τις διαδικασίες της δειγματοληψίας και της διακριτοποίησης, οπότε προκύπτει το τελικό ψηφιακό σήμα φωνής με το οποίο εργαζόμαστε.

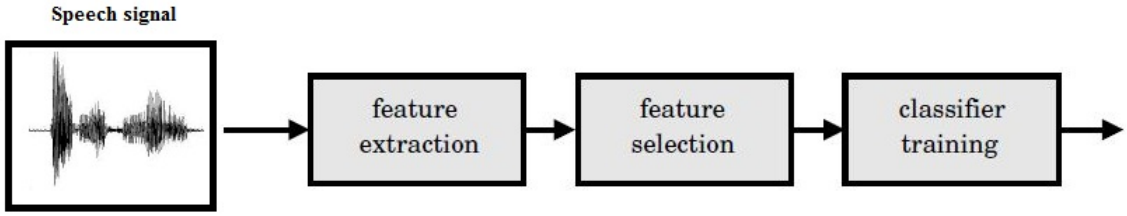
Η δομή της διπλωματικής αυτής εργασίας συμβαδίζει με την ακολουθία των βασικών βημάτων που προκύπτουν κατά την λύση του προβλήματος. Στο κεφάλαιο 2 περιγράφονται διάφορες κατηγορίες χαρακτηριστικών φωνής καθώς και ο τρόπος εξαγωγής τους, στο κεφάλαιο 3 μελετούνται βασικές μέθοδοι επιλογής χαρακτηριστικών, στο κεφάλαιο 4 αναλύονται ταξινομητές και ο τρόπος λειτουργίας τους, και στο κεφάλαιο 5 παρουσιάζονται αποτελέσματα από πειράματα αναγνώρισης στο σύστημα μας. Τέλος, στο κεφάλαιο 6 αναφέρουμε ορισμένα γενικά συμπεράσματα.

1.3 Η προσέγγιση της έρευνας μας και η πρόοδος πέραν της υπάρχουσας τεχνολογικής στάθμισης (state-of-the-art)

Στα πλαίσια της διπλωματικής αυτής, μελετάμε διάφορες κατηγορίες χαρακτηριστικών φωνής που μπορούν να διακρίνουν την συναισθηματική κατάσταση του ομιλητή. Εκτός από τις ευρύτατα χρησιμοποιημένες στη βιβλιογραφία, κατηγορίες των χαρακτηριστικών MFCC και



Σχήμα 1.3: Συναισθήματα στις τρείς διαστάσεις [10]



Σχήμα 1.4: Στάδια αναγνώρισης προτύπων

των χαρακτηριστικών προσωδίας (Pitch, Energy), δοκιμάζουμε και την ισχύ των χαρακτηριστικών του γλωττιδικού παλμού (Glottal flow features) καθώς και των AM-FM χαρακτηριστικών. Προσπαθούμε να συνδυάσουμε την διακριτική ικανότητα των διάφορων κατηγοριών χαρακτηριστικών, χρησιμοποιώντας αλγορίθμους επιλογής χαρακτηριστικών διάφορων τύπων. Για το τελικό στάδιο της αναγνώρισης, πειραματιζόμαστε κυρίως με τους ταξινομητές GMM και SVM και συγκρίνουμε τα αποτελέσματα. Μέρος της διπλωματικής δημοσιεύθηκε σε συνέδριο επεξεργασίας λόγου (LREC 2012) [11]. Παρακάτω συνοψίζονται τα πιο σημαντικά σημεία της διπλωματικής:

- Τόσο τα χαρακτηριστικά γλωττιδικού παλμού όσο και τα AM-FM, έδειξαν να σχετίζονται με το συναίσθημα στην φωνή, και ο συνδυασμός τους με τις κλασσικές κατηγορίες χαρακτηριστικών (MFCC, Prosody) βελτίωσε τα τελικά αποτελέσματα αναγνώρισης.

- Για την αποτελεσματική επιλογή χαρακτηριστικών, δοκιμάστηκαν και αλγόριθμοι φίλ-

τρου (F-score, MRMR), αλλά και αλγόριθμοι που εξαρτώνται από το ποσοστό αναγνώρισης του ταξινομητή (FSS, BSS). Στην δεύτερη κατηγορία, παρατηρήθηκαν καλύτερα αποτελέσματα με χρήση του αλγορίθμου BSS (Backward sequential selection) σε σχέση με τον FSS (Forward sequential selection). Τα καλύτερα συνολικά αποτελέσματα προέκυψαν με τον συνδυαστικό αλγόριθμο επιλογής δύο σταδίων (MRMR+BSS).

- Με χρήση των αλγορίθμων επιλογής, διαπιστώθηκε ότι η καθεμία από τις ψυχολογικές διαστάσεις των συναισθημάτων σχετίζεται περισσότερο με διαφορετικές ομάδες χαρακτηριστικών. Ακόμη σε άλλο πείραμα βρέθηκε ότι το καταλληλότερο σύνολο χαρακτηριστικών είναι διαφορετικό για κάθε ζεύγος συναισθημάτων.

- Στο πιο επιτυχημένο από τα πειράματα, κατασκευάστηκε ιεραρχικό σύστημα ταξινόμησης το οποίο στηρίζεται σε συγκρίσεις ανά δύο των συναισθημάτων με χρήση δυαδικού SVM. Με χρήση του αλγορίθμου επιλογής δύο σταδίων που προτείνουμε, επιλέγεται το βέλτιστο σύνολο χαρακτηριστικών για τον διαχωρισμό κάθε ζεύγους συναισθημάτων, και στην συνέχεια, η ταξινόμηση γίνεται σύμφωνα με το σχήμα του *majority vote*. Τα τελικά αποτελέσματα αναγνώρισης είναι συγκρίσιμα με τα καλύτερα ερευνητικά αποτελέσματα στην βιβλιογραφία για την συγκεκριμένη βάση δεδομένων [12], χρησιμοποιώντας όμως αρκετά μικρότερο αριθμό από εξαγόμενα ακουστικά χαρακτηριστικά.

- Τέλος, σε όλα τα πειράματα επιβεβαιώθηκε η υπεροχή των συστημάτων που χρησιμοποιούν την γνώση του φύλου του ομιλητή.

Κεφάλαιο 2

Μελέτη χαρακτηριστικών

Στην ενότητα αυτή θα παρουσιάσουμε τις διάφορες κατηγορίες ακουστικών χαρακτηριστικών που θα εξάγουμε από τα ηχογραφημένα στιγμιότυπα των ομιλητών, με σκοπό την αναγνώριση των συναισθημάτων τους. Προσπαθούμε να χρησιμοποιήσουμε τόσο κατηγορίες χαρακτηριστικών που σχετίζονται παραδοσιακά με εφαρμογές αναγνώρισης φωνής (Pitch, MFCC), όσο και με γενικά λιγότερο συνηθισμένες κατηγορίες που δείχνουν ωστόσο να σχετίζονται με το συναίσθημα στην φωνή (Glottal Flow, AM-FM). Επειδή τα περισσότερα χαρακτηριστικά σχετίζονται με τα μαθηματικά μοντέλα που χρησιμοποιούνται για την περιγραφή του μηχανισμού παραγωγής της φωνής, κρίνουμε χρήσιμο να παρουσιάσουμε αρχικά κάποια γενικά στοιχεία για αυτήν.

2.1 Εισαγωγή στην παραγωγή φωνής

Τα φωνητικά όργανα είναι οι πνεύμονες, ο αναπνευστικός σωλήνας, ο λάρυγγας (όπου υπάρχουν οι φωνητικές χορδές), ο λαιμός, η μύτη και το στόμα. Όλα μαζί αυτά τα όργανα σχηματίζουν ένα φωνητικό σωλήνα που ξεκινάει από τους πνεύμονες και καταλήγει στα χείλη (Σχήμα 2.2). Το τελευταίο τμήμα του σωλήνα αποτελείται από τον λαιμό και το στόμα και ονομάζεται φωνητική οδός.

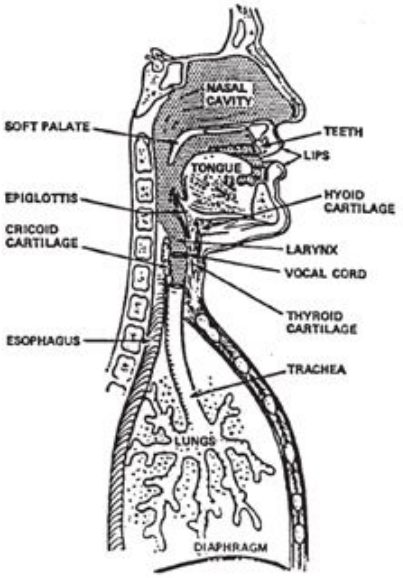
Η πηγή της ενέργειας για την παραγωγή της φωνής είναι ο αέρας που φεύγει από τους πνεύμονες κατά την εκπνοή. Ο αέρας αυτός διέρχεται μέσα από τον αναπνευστικό σωλήνα και φτάνει στον λάρυγγα όπου υπάρχουν οι φωνητικές χορδές. Μορφολογικά οι φωνητικές χορδές είναι ελαστικές πτυχές συνδέσμων που ενώνονται μεταξύ τους μόνο στο πρόσθιο τμήμα ενώ μπορούν να απομακρυνθούν στο οπίσθιο. Όταν οι φωνητικές χορδές είναι ανοιχτές, η ροή του αέρα περνάει μέσα από το άνοιγμα τους που έχει σχήμα 'V' και φτάνει στην φωνητική οδό, ενώ όταν είναι κλειστές αποτρέπουν την ροή του αέρα (Σχήμα 2.1). Όταν μιλάμε, οι φωνητικές χορδές ανοιγοκλείνουν γρήγορα, κατακερματίζοντας έτσι τη ροή του αέρα σε μια ακολουθία από διακοπτόμενες μάζες αέρα (puffs), η οποία μπορεί να ακουστεί ως ένας βόμβος (buzz), του οποίου η συχνότητα εξαρτάται από τον ρυθμό ταλάντωσης των φωνητικών χορδών. Παράλληλα κατά την διάρκεια της ομιλίας αλλάζει και το σχήμα της φωνητικής οδού με αποτέλεσμα να μεταβάλλονται και οι ακουστικές της ιδιότητες και να παράγονται έτσι οι διά-

Σχήμα 2.1: Κλείσιμο και άνοιγμα των φωνητικών χορδών [13]

φοροι ήχοι που επιθυμούμε (Σχήμα 2.6). Με άλλα λόγια η φωνητική οδός είναι ένας σωλήνας πολύπλοκου σχήματος ο οποίος μεταβάλλεται και λειτουργεί ως συντονιστής συχνοτήτων.

Οι περισσότεροι ήχοι παράγονται με τον τρόπο που περιγράψαμε παραπάνω. Υπάρχουν ωστόσο και κάποιες κατηγορίες ήχων οι οποίες παράγονται με ελαφρώς διαφορετικό τρόπο. Έτσι υπάρχουν οι τυρβώδεις ήχοι όπως το σύμφωνο /s/ οι οποίοι παράγονται με την δημιουργία κάποιας στένωσης κατά μήκος της φωνητικής οδού, και οι εκρηκτικοί ήχοι, όπως το σύμφωνο /p/, οι οποίοι παράγονται με στιγμιαία διακοπή της ροής του αέρα (φράζοντας με χείλια-δόντια) και την απότομη απελευθέρωση της στη συνέχεια.

Στα σχήματα 2.1-2.2 φαίνονται στιγμιότυπα από το άνοιγμα και το κλείσιμο των φωνητικών χορδών, καθώς και η ανατομία του συνολικού φωνητικού συστήματος στον άνθρωπο.



Σχήμα 2.2: Το ανθρώπινο σύστημα παραγωγής φωνής [13]

2.1.1 Μοντέλο πηγής-φίλτρου (source-filter model)

Το 1960 ο *Fant* [14] εισήγαγε ένα απλό μαθηματικό μοντέλο για να προσεγγίσει την διαδικασία παραγωγής της ομιλίας. Σύμφωνα με αυτό, ο μηχανισμός παραγωγής της φωνής μπορεί να μοντελοποιηθεί ως η έξοδος ενός γραμμικού φίλτρου το οποίο διεγείρεται από μια ανεξάρτητη από αυτό πηγή. Τον ρόλο του γραμμικού φίλτρου έχει η φωνητική οδός, ενώ τον ρόλο της πηγής έχει η ροή του αέρα που βγαίνει με κάποια συχνότητα από τις φωνητικές χορδές.

Οι βασικές υποθέσεις τις οποίες κάνει το μοντέλο αυτό είναι δύο: Α) Ότι η φωνητική οδός συμπεριφέρεται ως ένα γραμμικό φίλτρο όταν διεγείρεται από κάποια ροή αέρα. Β) Ότι οι φωνητικές χορδές και ο βόμβος που παράγουν είναι ανεξάρτητες από τον σχηματισμό της φωνητικής οδού.

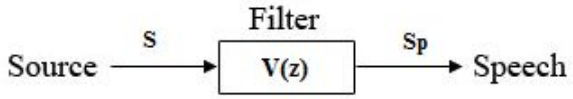
Όσον αφορά στην πρώτη παραδοχή, γνωρίζουμε ότι ένα σύστημα που αποτελείται από διαδοχικούς σωλήνες με στερεά τοιχώματα και δεν περιέχει απότομες και αιχμηρές απολήξεις κατά μήκος του, συμπεριφέρεται σαν γραμμικό ακουστικό σύστημα για ήχους λογικής έντασης. Η φωνητική οδός μπορεί να θεωρηθεί σε μεγάλο βαθμό ως ένα τέτοιο σύστημα από διαδοχικούς σωλήνες με στέρεα τοιχώματα (αν αγνοήσουμε την ελαστικότητα που υπάρχει στα μάγουλα). Μπορούμε έτσι να θεωρήσουμε χωρίς μεγάλο σφάλμα τον φωνητικό σωλήνα σαν ένα γραμμικό ακουστικό σύστημα.

Όσον αφορά στην δεύτερη παραδοχή, αυτό που ισχύει στην πράξη είναι ότι η φωνητική οδός, καθώς κινείται για να μεταβάλλει τον σχηματισμό της, επηρεάζει σε ένα μικρό βαθμό και τους μύες που ελέγχουν τις φωνητικές χορδές. Παράλληλα κάποιες φορές ένα μέρος της ενέργειας της ισχυρής πρώτης συχνότητας συντονισμού της φωνητικής οδού (1st formant) απορροφάται από τις φωνητικές χορδές επηρεάζοντας την ροή του αέρα που παράγουν. Το φαινόμενο αυτό γίνεται πιο έντονο όταν η πρώτη συχνότητα συντονισμού πλησιάζει στην θεμελιώδη συχνότητα των φωνητικών χορδών. Δεν είναι επομένως τελείως ανεξάρτητα στην πραγματικότητα. Ωστόσο ο βαθμός εξάρτησης μεταξύ της πηγής και της φωνητικής οδού δεν είναι και τόσο μεγάλος, και άρα μπορούμε προσεγγιστικά να δεχτούμε και την 2η παραδοχή.

Κάνοντας αυτές τις δύο παραδοχές προκύπτει το απλοποιημένο μοντέλο πηγής-φίλτρου. Το μοντέλο αυτό παρότι είναι προσεγγιστικό και απλό, καταφέρνει να προσομοιώσει τον γενικό μηχανισμό παραγωγής της φωνής και για αυτό έχει καθιερωθεί και χρησιμοποιείται μέχρι και σήμερα σε πολλές εφαρμογές με επιτυχία.

Σύμφωνα με το μοντέλο πηγής φίλτρου, αν s είναι το σήμα της ροής του αέρα που παράγεται από την πηγή, και V(z) η συνάρτηση μεταφοράς του γραμμικού φίλτρου της φωνητικής οδού, τότε το τελικό σήμα φωνής προκύπτει στην έξοδο με τον τρόπο που φαίνεται στο σχήμα 2.4.

Εάν μελετήσουμε το σήμα φωνής σε μικρό χρονικό διάστημα της τάξης των 25-30 msec, μπορούμε να θεωρήσουμε ότι το φίλτρο είναι και χρονικά αμετάβλητο. Έτσι δεδομένου ότι η φωνητική οδός είναι φίλτρο γραμμικό και χρονικά αμετάβλητο, στο πεδίο των συχνοτήτων ισχύει η σχέση 2.1:



Σχήμα 2.3: Γραμμικό μοντέλο παραγωγής φωνής

$$S_p(z) = S(z) \cdot V(z). \tag{2.1}$$

Συχνά προστίθεται σαν ξεχωριστό φίλτρο η επίδραση των χειλιών στην έξοδο, με συνάρτηση μεταφοράς R(z), οπότε προκύπτει η σχέση 2.2:

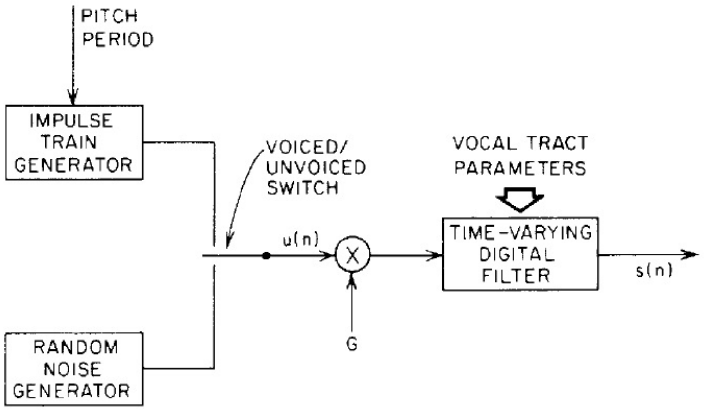
$$S_p(z) = S(z) \cdot V(z) \cdot R(z), \tag{2.2}$$

όπου η συνάρτηση μεταφοράς R(z) των χειλιών προσεγγίζεται συνήθως με μια γραμμική συνάρτηση σταθερής κλίσης.

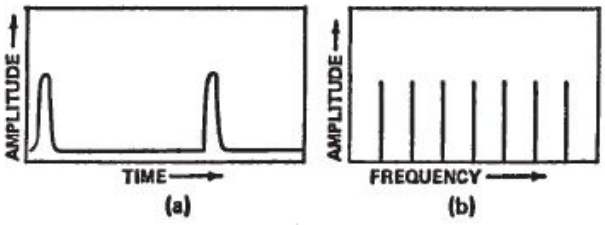
Σύμφωνα με το μοντέλο που περιγράψαμε, υπάρχουν δύο βασικοί τρόποι για την παραγωγή φωνής: Στην περίπτωση των έμφωνων ήχων, θεωρούμε ότι το σήμα της πηγής (δηλαδή ο αέρας που φεύγει από τις φωνητικές χορδές) έχει περιοδική μορφή. Το σήμα χαρακτηρίζεται από την θεμελιώδη συχνότητα του, περιέχει όμως και αρμονικές συχνότητες οι οποίες είναι ακέραια πολλαπλάσια της θεμελιώδους.

Στην περίπτωση των άφωνων ήχων, δεν υπάρχει ταλάντωση των φωνητικών χορδών, και το σήμα της πηγής που διεγείρει το φίλτρο θεωρείται ότι είναι θόρυβος. Το μοντέλο παραγωγής της φωνής για τις δύο περιπτώσεις ήχων φαίνεται στο σχήμα 2.5.

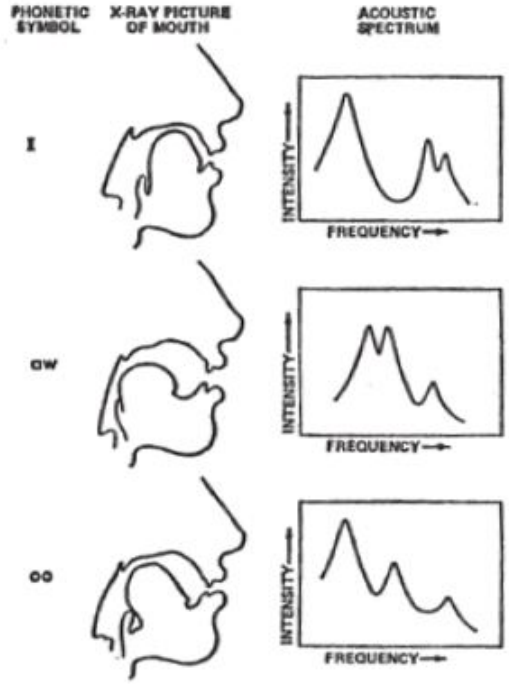
Τέλος δείχνουμε με μερικά σχήματα τι συμβαίνει στον χώρο των συχνοτήτων στα διάφορα στάδια παραγωγής ενός στιγμιότυπου φωνής, σύμφωνα πάντα με το μοντέλο πηγής-φίλτρου.



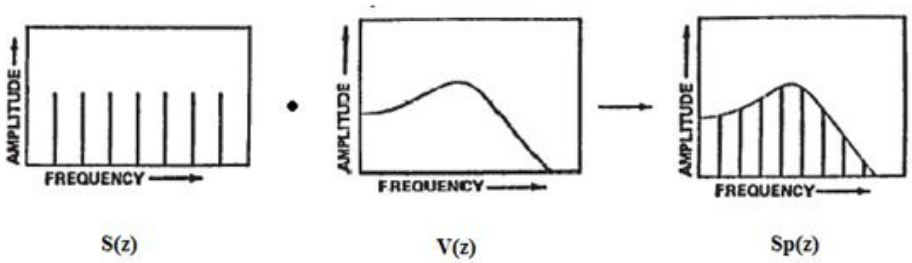
Σχήμα 2.4: Μοντέλο παραγωγής άφωνων και έμφωνων ήχων [15]



Σχήμα 2.5: Αναπαράσταση περιοδικής διέγερσης πηγής (α) στο πεδίο του χρόνου, (β) στο πεδίο των συχνοτήτων [13]



Σχήμα 2.6: Συναρτήσεις μεταφοράς του φίλτρου της φωνητικής οδού για διάφορους ήχους [13]



Σχήμα 2.7: Αναπαράσταση παραγωγής έμφωνου ήχου στο πεδίο της συχνότητας [13]

2.2 Χαρακτηριστικά προσωδίας

Με τον όρο "χαρακτηριστικά προσωδίας" εννοούμε τα χαρακτηριστικά εκείνα που σχετίζονται με την τονικότητα και την ένταση της φωνής. Τα κυριότερα είναι η θεμελιώδης συχνότητα (pitch) και η ενέργεια της φωνής. Τα χαρακτηριστικά προσωδίας έχουν χρησιμοποιηθεί ευρέως στην έρευνα της αναγνώρισης συναισθήματος μέσω φωνής με επιτυχία [16, 17, 18, 19].

2.2.1 Θεμελιώδης συχνότητα (Pitch)

Με τον όρο *pitch* εννοούμε την βασική συχνότητα με την οποία ένας ήχος γίνεται αντιληπτός. Το pitch αποτελεί θεμελιώδη ιδιότητα ενός ήχου η οποία καθορίζεται άμεσα από την συχνότητα των κυμάτων που τον παράγουν, και μετριέται σε Hertz (όπως και η συχνότητα). Εξαρτάται ωστόσο σε κάποιο βαθμό από τον τρόπο με τον οποίο το ανθρώπινο αυτί αντιλαμβάνεται τον ήχο και γι' αυτό έχει προβληματίσει αρκετά ο ακριβής ορισμός του.

Το pitch συνήθως συνδέεται με την απόσταση των συχνοτήτων των αρμονικών η οποία είναι ίση με την ελάχιστη-βασική συχνότητα F_0 του σήματος. Έχει καθιερωθεί λοιπόν να μετράμε σαν pitch την θεμελιώδη συχνότητα ενός σήματος, στην περίπτωση που αυτό έχει περιοδική ή περίπου περιοδική μορφή.

Η θεμελιώδης συχνότητα του σήματος έχει χρησιμοποιηθεί σαν χαρακτηριστικό σχεδόν σε όλες τις έρευνες για αναγνώριση συναισθήματος μέσω φωνής, γεγονός που είναι ενδεικτικό της άμεσης σχέσης του με την συναισθηματική κατάσταση του ομιλητή. Η ακριβής και αξιόπιστη εκτίμηση της θεμελιώδους συχνότητας θεωρείται δύσκολο πρόβλημα, και για την επίλυση του έχουν αναπτυχθεί αρκετοί αλγόριθμοι.

Στην δική μας έρευνα, για τον υπολογισμό του pitch χρησιμοποιήσαμε τον γνωστό αλγόριθμο RAPT (robust algorithm for pitch tracking) [20], ο οποίος θεωρείται ότι επιτυγχάνει αξιόπιστες εκτιμήσεις της θεμελιώδους συχνότητας ενός σήματος και μάλιστα χωρίς να επηρεάζεται ιδιαίτερα από την ύπαρξη θορύβου. Η λειτουργία του αλγορίθμου συνοψίζεται στα παρακάτω βήματα.

1. Κανονικοποιημένη συνάρτηση αυτοσυσχέτισης

Ο αλγόριθμος υπολογίζει μια εκτίμηση του F_0 που αντιστοιχεί στο κάθε πλαίσιο ανάλυσης (frame) στηριζόμενος στον υπολογισμό της κανονικοποιημένης αυτοσυσχέτισης του σήματος. Στο i-οστό πλαίσιο ανάλυσης, η κανονικοποιημένη συνάρτηση αυτοσυσχέτισης δίνεται από την παρακάτω σχέση:

$$f_{i,k} = \frac{\sum_{j=m}^{m+n-1} s_j s_{j+k}}{\sqrt{e_m e_{m+k}}}, \quad \mu\epsilon \;\; e_j = \sum_{i=j}^{j+n-1} s_i^2. \tag{2.3}$$

όπου n είναι το μήκος του πλαισίου, και $m = i * n$.

Όπως φαίνεται ισχύει ότι

$$-1 \leq f_{i,k} \leq 1 \tag{2.4}$$

ενώ οι τιμές της συνάρτησης τείνουν να είναι κοντά στο 1 σε θέσεις της μετατόπισης k που αντιστοιχούν σε ακέραια πολλαπλάσια της περιόδου, όταν πρόκειται για περιοδικά σήματα.

Αντίθετα όταν το σήμα μας είναι απεριοδικό και έχει τυχαία μορφή όπως π.χ. συμβαίνει με ένα τμήμα λευκού θορύβου, τότε οι τιμές της συνάρτησης είναι κοντά στο 0. Η βασική ιδέα λοιπόν είναι ότι τα περιοδικά σήματα έχουν περιοδικής μορφής αυτοσυσχέτιση και επομένως η θεμελιώδης συχνότητα αναζητείται μέσω αυτής.

2. 1ο πέρασμα σε αντίγραφο του σήματος που προκύπτει ύστερα από υπό-δειγματοληψία

Σαν πρώτο βήμα, ο αλγόριθμος χρησιμοποιεί ένα αντίγραφο του σήματος το οποίο είναι δειγματοληπτημένο σε αρκετά μικρότερο ρυθμό σε σχέση με το αρχικό ($Fs' << Fs$). Στη συνέχεια αφού υπολογίσει τις τιμές της κανονικοποιημένης αυτοσυσχέτισης για αυτό το σήμα ψάχνει για τοπικά μέγιστα τα οποία ξεπερνούν ένα δεδομένο κατώφλι (threshold).

3. 2ο πέρασμα στο αρχικό σήμα

Υπολογισμός της συνάρτησης αυτοσυσχέτισης του αρχικού σήματος, μόνο όμως στις περιοχές των k που επιλέχθηκαν από το προηγούμενο βήμα σαν θέσεις τοπικών μεγίστων. Με αυτό τον τρόπο υπολογίζουμε πιο λεπτομερείς εκτιμήσεις της πραγματικής θέσης των τοπικών μεγίστων αλλά και των πλατών τους. Όλες οι νέες θέσεις τοπικών μεγίστων που βρίσκονται από αυτό το 2ο και υψηλότερης ακρίβειας πέρασμα, αποτελούν τις υποψήφιες θεμελιώδεις συχνότητες F_0 για αυτό το πλαίσιο.

4. Επιλογή της σωστής θεμελιώδους συχνότητας $\mathbf{F}_0$ με χρήση δυναμικού προγραμματισμού

Στο τελικό αυτό βήμα εφαρμόζεται δυναμικός προγραμματισμός για να επιλεχθεί η καλύτερη υποψήφια F_0 για το κάθε πλαίσιο, συμπεριλαμβανομένης και της μηδενικής σε περίπτωση που το πλαίσιο κριθεί ως άφωνο. Η επιλογή της καλύτερης F_0 για το κάθε πλαίσιο γίνεται με χρήση κάποιων συναρτήσεων κόστους που ορίζουμε παρακάτω. Αξίζει να τονίσουμε ότι η επιλογή της κάθε F_0 εξαρτάται από τις επιλογές που γίνονται και στα γειτονικά πλαίσια. Έτσι ουσιαστικά ο αλγόριθμος ψάχνει για ένα βέλτιστο, από άποψη συνολικού κόστους, μονοπάτι από θεμελιώδεις συχνότητες.

Ορίζουμε το τοπικό κόστος για την επιλογή της j-οστής υποψήφιας περιόδου $\mathrm{L}_{i,j}$ για το πλαίσιο i ως:

$$d_{i,j} = 1 - C_{i,j} \cdot (1 - b \cdot L_{i,j}), \quad 1 \leq j \leq I_i, \tag{2.5}$$

ενώ για την μοναδική περίπτωση που θεωρούμε άφωνο το πλαίσιο, το τοπικό κόστος είναι:

$$d_{i,I_i} = V + \max_j (C_{i,j}), \tag{2.6}$$

όπου στις παραπάνω σχέσεις με $C_{i,j}$ συμβολίζουμε την τιμή του j-οστού τοπικού μέγιστου της συνάρτησης αυτοσυσχέτισης $f_{i,k}$ στο πλαίσιο i, και με I_i τον αριθμό των υποψήφιων συχνοτήτων που προέκυψαν έπειτα και από το 2ο πέρασμα για το πλαίσιο i. Τα b, V είναι σταθερές.

Όπως φαίνεται από την σχέση 2.5 το τοπικό κόστος για την επιλογή της j υποψήφιας συχνότητας είναι μικρότερο για μεγάλες τιμές του πλάτους $C_{i,j}$ στην συνάρτηση αυτοσυσχέτισης, ενώ η τιμή b μας επιτρέπει να ρυθμίσουμε τον βαθμό στον οποίο μικρότερες τιμές περιόδων θα προτιμώνται σε σχέση με μεγαλύτερες. Στην σχέση 2.6 φαίνεται ότι το κόστος χαρακτηρισμού ενός πλαισίου ως άφωνο είναι μεγάλο όταν οι υποψήφιες συχνότητες έχουν μεγάλα πλάτη στην συνάρτηση αυτοσυσχέτισης. Η σταθερά V αποτελεί κατώφλι το οποίο μπορούμε να το ρυθμίζουμε ανάλογα με το σήμα μας.

Ορίζουμε στη συνέχεια τις συναρτήσεις κόστους μετάβασης. Κόστος μετάβασης από έμφωνο σε έμφωνο πλαίσιο:

$$\delta_{i,j,k} = F \times \min\{\xi_{i,j,k}, (J + |\xi_{i,j,k} - \ln(2)|)\}, \tag{2.7}$$

$$\xi_{i,j,k} = |ln\frac{L_{i,j}}{L_{i-1,k}}|, \quad 1 \leq j \leq I_i, \quad 1 \leq k \leq I_{i-1}, \tag{2.8}$$

όπου F και J σταθερές, και $\xi_{i,j,k}$ ο λογαριθμικός λόγος της j-οστής υποψήφιας περιόδου για το πλαίσιο i, προς την k-οστή υποψήφια περίοδο για το προηγούμενο πλαίσιο, $i-1$. Όπως φαίνεται από τις προηγούμενες σχέσεις, το κόστος μετάβασης αυξάνεται όταν αυξάνεται και η διαφορά μεταξύ των θεμελιωδών συχνοτήτων των γειτονικών πλαισίων. Η σταθερά J μας επιτρέπει να ρυθμίζουμε το κόστος του άλματος οκτάβας.

Κόστος μετάβασης από άφωνο σε άφωνο πλαίσιο:

$$\delta_{i,I_i,I_{i-1}} = 0 \tag{2.9}$$

Κόστος μετάβασης από έμφωνο σε άφωνο πλαίσιο:

$$\delta_{i,I_i,k} = V_a + V_b \cdot S_i + V_c \cdot rr_i \tag{2.10}$$

Κόστος μετάβασης από άφωνο σε έμφωνο πλαίσιο:

$$\delta_{i,j,I_{i-1}} = V_a + V_b \cdot S_i + V_c/rr_i, \tag{2.11}$$

όπου

$$S_i = \frac{0.2}{D_I(i,i-1) - 0.8}, \quad rr_i = \frac{rms(i)}{rms(i-1)}, \tag{2.12}$$

και τα V_a, V_b, V_c είναι σταθερές. Το S_i είναι αντιστρόφως ανάλογο της φασματικής απόστασης Itakura του σήματος (D_I) υπολογισμένης στο σύνορο των πλαισίων i και $i-1$. Εάν στο i-οστό πλαίσιο παρατηρείται αύξηση του πλάτους του σήματος σε σχέση με το προηγούμενο πλαίσιο, τότε $rr_i > 1$, αλλιώς $rr_i < 1$. Οι δυο αυτές ποσότητες χρησιμοποιούνται για να μειώσουν το κόστος αλλαγής φωνητικής κατάστασης (από έμφωνη σε άφωνη ή αντίστροφα) στις περιπτώσεις που το φάσμα ή το πλάτος του σήματος μεταβληθούν σχετικά γρήγορα και απότομα.

Μπορούμε τώρα να ορίσουμε την αναδρομή που υπολογίζει την συνάρτηση τελικού κόστους για κάθε πλαίσιο ανάλυσης i.

$$D_{i,j} = d_{i,j} + \min_{k \in I_{i-1}} (D_{i-1,k} + \delta_{i,j,k}) \tag{2.13}$$

με αρχικές συνθήκες $D_{0,j} = 0$

Στην αναδρομή φαίνεται ότι το τελικό κόστος επιλογής της j-οστής υποψήφιας F_0 για το i-οστό πλαίσιο εξαρτάται από το τοπικό κόστος και από το κόστος μετάβασης από την προηγούμενη κατάσταση. Στόχος είναι η ελαχιστοποίηση του κόστους $D_{i,j}$ για κάθε πλαίσιο i.

Για κάθε κατάσταση και σε κάθε πλαίσιο κρατάμε τους δείκτες $q_{i,j} = k_{min}$ έτσι ώστε να μπορέσουμε μετά να ανακτήσουμε το βέλτιστο μονοπάτι από θεμελιώδεις συχνότητες, δηλαδή την τελική εκτίμηση του F_0 ($F_{0i} = F_s$ / $L_{i,j}$) για κάθε ένα από τα πλαίσια.

Ο αλγόριθμος RAPT έχει εφαρμοσθεί με επιτυχία σε αρκετές περιοχές επεξεργασίας φωνής όπως είναι η αναγνώριση και η σύνθεση φωνής. Αν και έχει αυξημένο υπολογιστικό κόστος σε σχέση με άλλες απλούστερες προσεγγίσεις (amdf method, cepstrum method) κερδίζει σε αξιοπιστία και σε ακρίβεια και γι΄ αυτό προτιμήθηκε στην έρευνα μας.

Pitch και συναίσθημα

Όσον αφορά στην συμπεριφορά του pitch στα διάφορα συναισθήματα, οι περισσότερες έρευνες έχουν καταλήξει στο ότι το pitch έχει μεγαλύτερη μέση τιμή και εύρος τιμών στη χαρά και στο θυμό, ενώ μικρότερες τιμές στην λύπη και την αποστροφή. Στο θυμό έχουν βρεθεί απότομες διακυμάνσεις του pitch, ενώ στη χαρά έχει πιο ομαλή πορεία. Φθίνουσα πορεία των τιμών του παρατηρείται στα συναισθήματα της λύπης και της αποστροφής [21]. Τέλος, για τον φόβο έχουν παρατηρηθεί πολύ υψηλές τιμές του pitch με μεγάλο εύρος και κανονική διακύμανση [22].

2.2.2 Ένταση-ενέργεια ομιλίας

Για την ενέργεια της ομιλίας,η οποία σχετίζεται άμεσα με την ένταση της φωνής, έχουν γίνει πολλές έρευνες [18, 23, 24, 19]. Σε μία μάλιστα [24] έχει βρεθεί ότι χαρακτηριστικά της ενέργειας του σήματος, όπως η μέγιστη τιμή, η διάρκεια του μέγιστου επιπέδου της και η κλίση της, ειναι πολύ πιο αποδοτικά σε σχέση με άλλα χαρακτηριστικά του pitch και των formants, ειδικά όσον αφορά στη γυναικεία φωνή [23].

Στα δικά μας πειράματα, υπολογίζουμε την ενέργεια του σήματος ανά πλαίσιο ανάλυσης, και στη συνέχεια χρησιμοποιούμε σαν χαρακτηριστικά κυρίως την μέση τιμή, το μέγιστο, και το εύρος τιμών της για κάθε εκφωνημένη πρόταση.

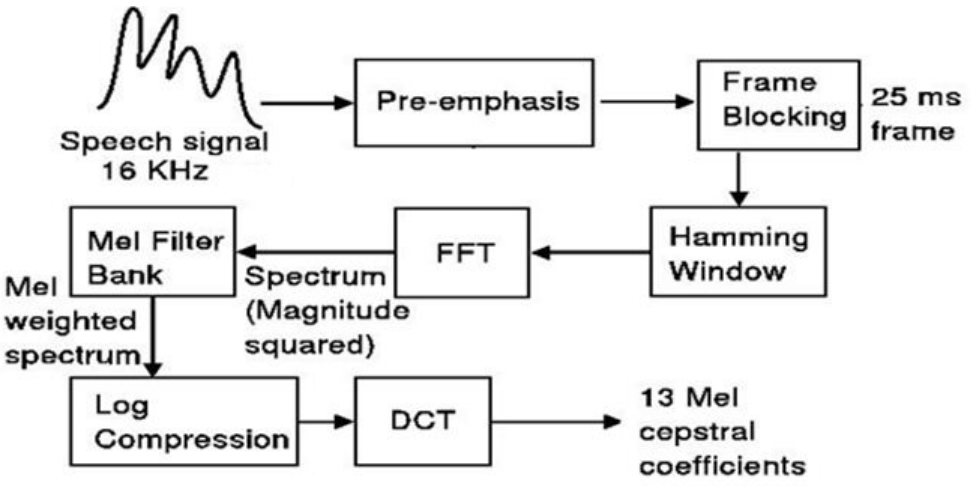
2.3 Χαρακτηριστικά φάσματος - Χαρακτηριστικά φωνητικής οδού

Τα χαρακτηριστικά φάσματος προκύπτουν έπειτα από επεξεργασία του σήματος στο πεδίο των συχνοτήτων, και αποτελούν την πιο συνηθισμένη κατηγορία χαρακτηριστικών. Στην επεξεργασία φωνής, συχνά εξάγονται χαρακτηριστικά φάσματος τα οποία σχετίζονται με τον σχηματισμό της φωνητικής οδού κατά την διάρκεια της ομιλίας. Στην έρευνα αναγνώρισης συναισθήματος έχουν χρησιμοποιηθεί σαν χαρακτηριστικά οι συχνότητες των Formants [25, 26, 27, 22], οι συντελεστές LPC [4], οι οποίοι περιέχουν πληροφορία για την περιβάλλουσα του φάσματος της φωνής, και τέλος, οι συντελεστές MFCC. Στα δικά μας πειράματα θα δοκιμάσουμε μόνο την τελευταία κατηγορία χαρακτηριστικών, η οποία είναι και η πιο επιτυχημένη για εφαρμογές επεξεργασίας φωνής.

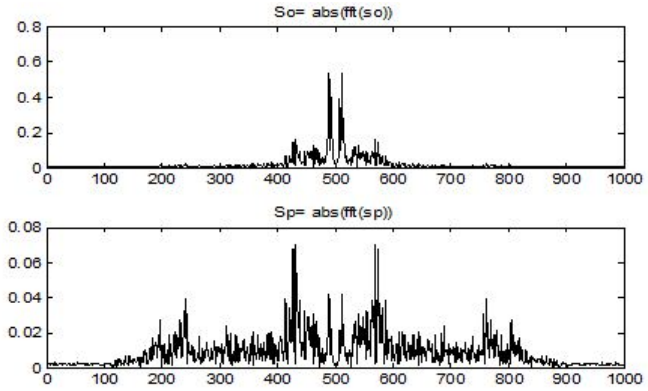
2.3.1 MFCCs (Mel frequency cepstral coefficients)

Οι MFCCs αποτελούν το περισσότερο διαδεδομένο σύνολο χαρακτηριστικών σε εφαρμογές αναγνώρισης φωνής. Αποτελούν μια συμπαγή αναπαράσταση του φάσματος του σήματος φωνής και περιέχουν μεικτή πληροφορία καθώς σχετίζονται τόσο με τον ομιλητή και την κατάσταση στην οποία βρίσκεται όσο με τα λεγόμενα του. Στην αναγνώριση συναισθήματος έχουν χρησιμοποιηθεί σε αρκετές έρευνες και πολλές φορές δίνουν ικανοποιητικά αποτελέσματα [28, 29, 30]. Δεν θα μπορούσαμε λοιπόν να τα παραλείψουμε από την έρευνα μας. Στην συνέχεια δίνουμε μια περιγραφή των συντελεστών αυτών και της διαδικασίας εξαγωγής τους από το σήμα. Στο σχήμα 2.8 φαίνεται η διαδικασία εξαγωγής των χαρακτηριστικών MFCC από ένα δείγμα φωνής.

Οι συντελεστές MFCC είναι στην ουσία μια ομάδα συντελεστών cepstrum που εξάγονται



Σχήμα 2.8: Διαδικασία εξαγωγής χαρακτηριστικών $MFCC$



Σχήμα 2.9: Σχήμα προενίσχυσης του σήματος φωνής

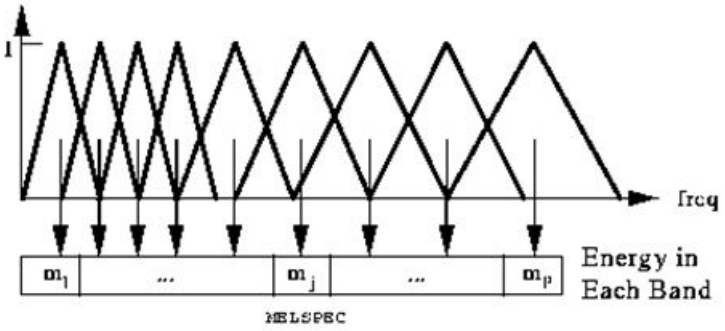
μετά από ανάλυση του σήματος με μια ειδικά σχεδιασμένη συστοιχία φίλτρων (Mel Filter Bank). Αρχικά περνάμε το σήμα μας $s_0[n]$ από ένα σύστημα προέμφασης δίνοντας στην έξοδο το $s_p[n]$. Σκοπός της προέμφασης είναι να ενισχύσει τις υψηλές συχνότητες σε σχέση με τις χαμηλότερες, πετυχαίνοντας έτσι ένα spectral flattening στο σήμα μας και βελτιώνοντας το σηματοθορυβικό του λόγο (SNR).

Στη συνέχεια χωρίζουμε το σήμα μας σε επικαλυπτόμενα πλαίσια διάρκειας T= 25 msec με επικάλυψη $\mathrm{T}_{overlap} = 10msec$, και παραθυρώνουμε το κάθε πλαίσιο με παράθυρο Hamming. Η παραθύρωση με Hamming window γίνεται για να εξομαλυνθούν οι ασυνέχειες στα άκρα των πλαισίων μετά την εφαρμογή του γρήγορου μετασχηματισμού Fourier (FFT). Αφού παραθυρώσουμε κάθε πλαίσιο, υπολογίζουμε τον αντίστοιχο διακριτό μετασχηματισμό Fourier (DFT) $N = 512$ σημείων. Μένει να φιλτράρουμε το μετασχηματισμένο σήμα με την συστοιχία φίλτρων που προαναφέραμε. Η ανάλυση γίνεται με συστοιχία φίλτρων σε κλίμακα Mel. Η συστοιχία φίλτρων αποτελείται από Q τριγωνικά φίλτρα H_j ,j=1,...,Q, των οποίων οι κεντρικές συχνότητες f_c είναι ισοκατανεμημένες στην κλίμακα Mel και συνδέονται με τις αντίστοιχες γραμμικές σύμφωνα με την σχέση 2.14.

$$f_c^{(j)} = 2595 \cdot \log(1 + \frac{f^{(j)}}{700}) \tag{2.14}$$

Η κλίμακα συχνοτήτων Mel είναι εμπνευσμένη από ψυχοακουστικές μελέτες και προσπαθεί να προσομοιώσει την ανάλυση συχνοτήτων του ανθρώπινου αυτιού [19]. Στο σχήμα 2.11 φαίνεται η Mel Filter Bank.

Όπως φαίνεται από το σχήμα, η ανάλυση του κάθε πλαισίου φωνής με την συστοιχία φίλτρων Mel μας δίνει σημαντική πληροφορία για την ενέργεια του συγκεκριμένου πλαισίου στις συγκεκριμένες κεντρικές συχνότητες της κλίμακας Mel. Έτσι, έπειτα από την εφαρμογή του φίλτρου έχουμε βρει τους συντελεστές E_j, $j = 1,...,Q$ καθένας από τους οποίους αντιπροσωπεύει την ενέργεια που συγκεντρώνει το σήμα μας στην περιοχή της αντίστοιχης συχνότητας $f_c^{(j)}$.



Σχήμα 2.10: Συστοιχία φίλτρων κλίμακας Mel [31]

Στη συνέχεια υπολογίζουμε τους ακόλουθους συντελεστές:

$$G_j = \log(E_j), \quad j = 1, .., Q \tag{2.15}$$

Τέλος εφαρμόζοντας διακριτό μετασχηματισμό συνημιτόνου (DCT) στους G_j συντελεστές, λαμβάνουμε τα ζητούμενα ακουστικά χαρακτηριστικά, ή αλλιώς τους MFCC (Mel Frequency Cepstrum Coefficients) από την σχέση:

$$C(n) = \sum_{j=1}^{Q} G(j) \cos(n(j - \frac{1}{2})\frac{\pi}{Q}) \tag{2.16}$$

Η εφαρμογή του μετασχηματισμού DCT γίνεται για να μειωθεί η επικάλυψη πληροφορίας μεταξύ των γειτονικών ζωνών συχνοτήτων που έχουμε επιλέξει. Για κάθε πλαίσιο λοιπόν, εξάγουμε τους 13 συντελεστές $MFCC$ οι οποίοι θα το αντιπροσωπεύουν στην παρακάτω μελέτη μας. Η σημαντικότητα της πληροφορίας που περιέχουν θα φανεί στη συνέχεια.

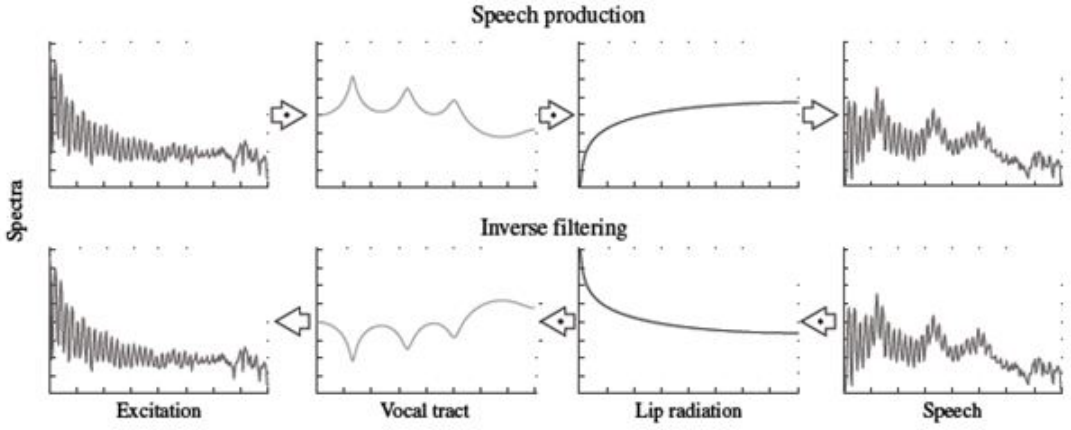
Έχοντας εξάγει τους συντελεστές $MFCC$ για κάθε πλαίσιο του δείγματος φωνής, μπορούμε πλέον να υπολογίσουμε και τις χρονικές παραγώγους 1ης και 2ης τάξης και να τις θεωρήσουμε ως επιπλέον χαρακτηριστικά.

2.4 Χαρακτηριστικά γλωττιδικού παλμού (*Glottal flow features*)

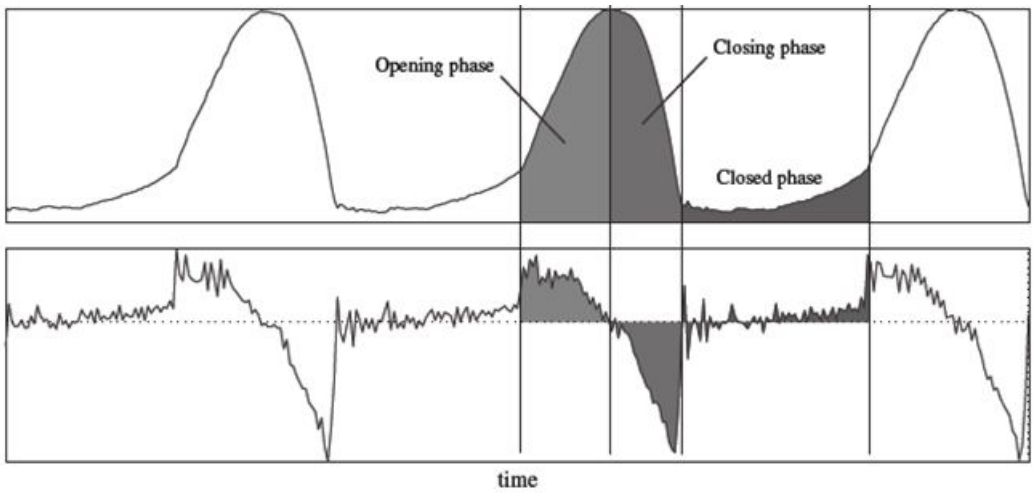
Όπως αναφέραμε και προηγουμένως, οι συντελεστές $MFCC$ αποτελούν το πιο διαδεδομένο σύνολο χαρακτηριστικών για εφαρμογές αναγνώρισης φωνής. Παρόλο που περιέχουν μεικτή πληροφορία για το φάσμα συχνοτήτων του σήματος μας, σχετίζονται κυρίως με τον σχηματισμό της φωνητικής οδού του ομιλητή και με την φράση που εκφωνείται. Χάνουν έτσι ένα σημαντικό μέρος της πληροφορίας της φωνής, αυτό που σχετίζεται με τον γλωττιδικό παλμό της πηγής. Τα χαρακτηριστικά προσωδίας από την άλλη, περιέχουν πολύ γενική πληροφορία για τον γλωττιδικό παλμό. Πιο συγκεκριμένα, η θεμελιώδης συχνότητα αναφέρεται στην απόσταση μεταξύ διαδοχικών παλμών του γλωττιδικού σήματος της πηγής, ενώ η ενέργεια μας δίνει πληροφορία για το μέσο πλάτος τους. Δεν παρέχεται όμως καμιά πληροφορία για το σχήμα και την μορφή των παλμών αυτών, χαρακτηριστικά τα οποία παίζουν καθοριστικό ρόλο για την "ποιότητα" της φωνής που ακούμε και συνδέονται άμεσα με την συναισθηματική κατάσταση στην οποία βρίσκεται ο ομιλητής. Θα επιχειρήσουμε λοιπόν να ανακτήσουμε από το σήμα φωνής μας τον γλωττιδικό παλμό, που αποτελεί το άλλο ανεξάρτητο μέρος του μοντέλου πηγής - φίλτρου, και να αναλύσουμε στη συνέχεια διάφορα χαρακτηριστικά του. Η διαδικασία υπολογισμού του γλωττιδικού παλμού ή αλλιώς του σήματος πηγής, ονομάζεται αντίστροφο φιλτράρισμα (inverse filtering). Στη συνέχεια θα παρουσιάσουμε συνοπτικά τις πιο γνωστές τεχνικές που έχουν αναπτυχθεί για αντίστροφο φιλτράρισμα και θα αναλύσουμε την μέθοδο που επιλέξαμε να χρησιμοποιήσουμε στην έρευνα μας.

Η βασική ιδέα των μεθόδων υπολογισμού του σήματος της πηγής είναι η εκτίμηση του γραμμικού φίλτρου της φωνητικής οδού. Αν γνωρίζαμε την εξίσωση του φίλτρου, τότε στηριζόμενοι στο μοντέλο πηγής-φίλτρου, για να αποκτήσουμε το σήμα πηγής, θα περνούσαμε απλά το σήμα φωνής από το αντίστροφο φίλτρο. Η έξοδος θα είναι το σήμα πηγής. Όσο καλύτερη είναι η εκτίμηση του φίλτρου της φωνητικής οδού, τόσο καλύτερη θα είναι και αυτή του σήματος πηγής. Στο σχήμα 2.11 φαίνονται η διαδικασία παραγωγής της φωνής και η αντίστροφη διαδικασία για την απόκτηση του φάσματος του σήματος πηγής: Το σήμα φωνής περνάει από έναν ολοκληρωτή για να απαλείψει την επίδραση των χειλιών (lips radiation effect) και στη συνέχεια περνάει από ένα φίλτρο αντίστροφο αυτού της φωνητικής οδού, για να μας δώσει το φάσμα που θέλουμε.

Στην περίπτωση έμφωνου ήχου, η μορφή του παλμού της πηγής που θα πάρουμε στην έξοδο θα είναι παρόμοια με αυτή στο σχήμα 2.12. Διακρίνουμε σε κάθε περίοδο τρείς κύριες φάσεις: Το άνοιγμα των φωνητικών χορδών, το κλείσιμο τους, και το διάστημα που παραμένουν κλειστές μέχρι να ανοίξουν πάλι από την αρχή.



Σχήμα 2.11: Στα σχήματα φαίνεται το λογαριθμικό φάσμα των σημάτων και φίλτρων στις διάφορες φάσεις της παραγωγής φωνής στην 1η σειρά, και της διαδικασίας αντίστροφου φιλτραρίσματος στην 2η [32]



Σχήμα 2.12: Μορφή του γλωττιδικού παλμού και της παραγώγου του [32]

2.4.1 Μέθοδος γραμμικής πρόβλεψης - LPC (linear predictive coding) για αντίστροφο φιλτράρισμα.

Η μέθοδος αυτή είναι από τις πρώτες που χρησιμοποιήθηκαν για τον υπολογισμό του φίλτρου της φωνητικής οδού. Στηρίζεται στην ευρύτατα διαδεδομένη μοντελοποίηση του σήματος φωνής με την μέθοδο της γραμμικής πρόβλεψης [33]. Σύμφωνα με το μοντέλο γραμμικής πρόβλεψης, τα διαδοχικά στιγμιότυπα του σήματος φωνής δεν είναι ανεξάρτητα μεταξύ τους αλλά

σχετίζονται σε σημαντικό βαθμό. Μπορούμε έτσι να προσεγγίσουμε την τιμή του σήματος φωνής σε μια χρονική στιγμή με ένα γραμμικό συνδυασμό των τιμών των p προηγούμενων δειγμάτων. Έτσι έχουμε ότι:

$$\hat{s}[n] = \sum_{k=1}^{p} a_k s[n-k]. \tag{2.17}$$

Ο προσδιορισμός των παραμέτρων a_k γίνεται έτσι ώστε να ελαχιστοποιηθεί η ενέργεια του σφάλματος από το πραγματικό σήμα φωνής $s[n]$, το οποίο θεωρούμε ότι παράγεται σύμφωνα με το μοντέλο πηγής-φίλτρου. Το φίλτρο το προσεγγίζουμε με μια συνάρτηση μεταφοράς που έχει μόνο πόλους ($all-pole\ model$). Έτσι έχουμε για τη συνάρτηση μεταφοράς του φίλτρου:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{S(z)}{U(z)}, \tag{2.18}$$

όπου $S(z)$ είναι ο μετασχηματισμός -Z του τελικού σήματος φωνής και $U(z)$ ο μετασχηματισμός της διέγερσης της πηγής. Σύμφωνα με το μοντέλο πηγής-φίλτρου, η διέγερση $u[n]$ θεωρείται περιοδική παλμοσειρά για τους έμφωνους ήχους και θόρυβος για τους άφωνους. Από την προηγούμενη σχέση προκύπτει για το σήμα φωνής ότι:

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + Gu[n], \tag{2.19}$$

ενώ για το σφάλμα και την ενέργεια του έχουμε ότι:

$$e[n] = s[n] - \hat{s}[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k] \tag{2.20}$$

$$E = \sum_{n} (s[n] - \hat{s}[n])^2 \tag{2.21}$$

Για τον υπολογισμό των συντελεστών a_i απαιτούμε να ελαχιστοποιηθεί η ενέργεια του λάθους (για το πλαίσιο του σήματος το οποίο εξετάζουμε):

$$\frac{\partial E}{\partial a_i} = 0, \quad i = 1, 2, .., p, \tag{2.22}$$

οπότε προκύπτουν οι παρακάτω εξισώσεις:

$$\sum_{n} s[n-i]s[n] = \sum_{k=1}^{p} a_k \sum_{n} s[n-i]s[n-k], \quad i = 1, 2, .., p \tag{2.23}$$

οι οποίες αν ορίσουμε την συνάρτηση συσχέτισης ϕ,

$$\phi[i,k] = \sum_{n} s[n-i]s[n-k] \tag{2.24}$$

γράφονται στην μορφή:

$$\sum_{k=1}^{p} a_k \phi[i,k] = \phi[i,0] \quad , i = 1, 2, .., p \tag{2.25}$$

Ανάλογα με τα όρια άθροισης που ορίζουμε στον τύπο της ενέργειας σφάλματος, προκύπτουν δύο διαφορετικές προσεγγίσεις του προβλήματος. Στην 1η περίπτωση προσπαθούμε να ελαχιστοποιήσουμε την ενέργεια του λάθους σε όλο το εύρος του χρόνου, από $-\infty$ έως το $+\infty$. Έτσι όταν εξετάζουμε ένα μεμονωμένο πλαίσιο του σήματος φωνής, θεωρούμε ότι το σήμα φωνής εκτείνεται σε όλο το χρόνο και απλά λαμβάνει μηδενικές τιμές έξω από το διάστημα που εξετάζουμε. Αν λοιπόν το πλαίσιο είναι μήκους N, τότε θεωρούμε ότι το σήμα μας παίρνει μη μηδενικές τιμές στο διάστημα 0 έως $N-1$, ενώ το σήμα λάθους θα είναι μη μηδενικό στο διάστημα 0 έως $N-1+p$.

Για την ενέργεια του λάθους θα έχουμε λοιπόν ότι

$$E = \sum_{n=-\infty}^{\infty} e^2[n] = \sum_{n=0}^{N-1+p} e^2[n] \tag{2.26}$$

ενώ για την συνάρτηση συσχέτισης ϕ προκύπτει:

$$\phi[i,k] = \sum_{n=0}^{N-1+p} s[n-i]s[n-k] = \sum_{n=0}^{N-1-|i-k|} s[n]s[n-|i-k|] = R_s[|i-k|], \tag{2.27}$$

όπου R_s η συνάρτηση αυτοσυσχέτισης του σήματος s. Ο συμμετρικός πίνακας της συνάρτησης ϕ τώρα γίνεται $Toeplitz$. Οι συμμετρικές ιδιότητες του πίνακα ϕ επιτρέπουν την ύπαρξη αποδοτικού αλγορίθμου (αλγόριθμος $Levinson-Durbin$) για την επίλυση των p κανονικών εξισώσεων και την εύρεση των παραμέτρων a_i. Η μέθοδος αυτή ονομάζεται μέθοδος της αυτοσυσχέτισης (autocorrelation method).

Στην 2η περίπτωση, το λάθος πρόβλεψης υπολογίζεται μόνο στο συγκεκριμένο διάστημα που εξετάζουμε το σήμα μας, και οι τιμές του σήματος που χρειάζονται ανήκουν στο διάστημα $-p \leq n \leq N-1$. Έτσι έχουμε:

$$E = \sum_{n=0}^{N-1} e^2[n] \tag{2.28}$$

και για την συνάρτηση συσχέτισης ϕ,

$$\phi[i,k] = \sum_{n=0}^{N-1} s[n-i]s[n-k] \quad , 0 \leq i,k \leq p \tag{2.29}$$

Ο πίνακας ϕ σε αυτή την περίπτωση δεν είναι Toeplitz αλλά απλά συμμετρικός, και έτσι χρειάζεται επίλυση p κανονικών εξισώσεων με κλασσικές μεθόδους που κοστίζουν περισσότερο υπολογιστικά. Η μέθοδος αυτή ωστόσο παράγει πιο ακριβείς εκτιμήσεις του φίλτρου της φωνητικής οδού και μπορεί να εφαρμοστεί και σε μικρότερα διαστήματα με επιτυχία. Η μέθοδος αυτή ονομάζεται μέθοδος της συμμεταβλητότητας (covariance method).

Στην συνέχεια θα αναφέρουμε συνοπτικά τις διάφορες μεθόδους που έχουν επικρατήσει για την διαδικασία του αντίστροφου φιλτραρίσματος.

1. Μέθοδος αυτοσυσχέτισης στο αρχικό σήμα

Χωρίζουμε το σήμα μας σε πλαίσια βραχείας διάρκειας των 25-30 msec, έτσι ώστε το γραμμικό φίλτρο της φωνητικής οδού να μπορεί να θεωρηθεί και χρονικά αμετάβλητο. Στη συνέχεια σε κάθε πλαίσιο υπολογίζουμε το μοντέλο γραμμικής πρόβλεψης με την μέθοδο αυτοσυσχέτισης και υπολογίζουμε τους συντελεστές a_i του φίλτρου. Τέλος δίνουμε το αρχικό σήμα φωνής σαν είσοδο στο φίλτρο που υπολογίσαμε και παίρνουμε στην έξοδο την εκτίμηση της γλωττιδικής παλμοσειράς.

Το βασικό μειονέκτημα της μεθόδου αυτοσυσχέτισης είναι ότι, ειδικά για έμφωνους ήχους, συχνά χάνει σε ακρίβεια. Αυτό οφείλεται στην απλοποιητική παραδοχή ότι το σήμα μας είναι μηδενικό έξω από το πλαίσιο που εξετάζουμε. Έτσι τα πρώτα p και τα τελευταία p δείγματα του λάθους πρόβλεψης μπορούν να πάρουν μεγάλες τιμές, αφού προσπαθούμε ουσιαστικά να προβλέψουμε μη μηδενικά από μηδενικά δείγματα και το αντίστροφο. Το λάθος πρόβλεψης μάλιστα θεωρητικά δεν μπορεί ποτέ να είναι μηδενικό.

2. Εντοπισμός κλειστής φάσης φωνητικών χορδών και εφαρμογή μεθόδου συμμεταβλητότητας

Όπως είχαμε αναφέρει προηγουμένως στις παραδοχές του μοντέλου πηγής-φίλτρου, η πηγή και το φίλτρο δεν είναι τελείως ανεξάρτητα μεταξύ τους. Η μέθοδος αυτή προσπαθεί να εντοπίσει τα διαστήματα στα οποία οι φωνητικές χορδές είναι κλειστές, αφού στα διαστήματα αυτά η υπόθεση της ανεξαρτησίας και της έλλειψης αλληλεπίδρασης μεταξύ πηγής και φίλτρου ισχύει σε μεγάλο βαθμό. Αφού εντοπίσει τα διαστήματα αυτά εφαρμόζει τον αλγόριθμο γραμμικής πρόβλεψης με την μέθοδο συμμεταβλητότητας η οποία πετυχαίνει περισσότερο ακριβείς εκτιμήσεις του φίλτρου της φωνητικής οδού σε σχέση με την μέθοδο αυτοσυσυσχέτισης.

Ένα από τα προβλήματα της μεθόδου αυτής είναι ότι συχνά η κλειστή φάση των φωνητικών χορδών διαρκεί για πολύ μικρό χρονικό διάστημα, με αποτέλεσμα να έχουμε λίγα δεδομένα για να υπολογίσουμε με ακρίβεια το μοντέλο γραμμικής πρόβλεψης με την μέθοδο συμμεταβλητότητας. Επιπλέον η επιτυχία του αλγορίθμου εξαρτάται από την ακριβή εύρεση της περιοχής κλειστής φάσης. Αν μάλιστα λάβουμε υπόψιν μας το μικρό μήκος της, ένα μικρό λάθος στην επιλογή της περιοχής θα δημιουργούσε σημαντικά προβλήματα στην σωστή εκτίμηση του φίλτρου.

3. Μέθοδος IAIF με χρήση Discrete All-Pole Modeling

Όπως αναφέραμε και προηγουμένως η μέθοδος της αυτοσυσχέτισης παρουσιάζει προβλήματα στην ακριβή εύρεση της περιβάλλουσας του φίλτρου, ειδικά όταν το πλαίσιο φωνής αντιστοιχεί σε έμφωνο ήχο. Μπορούμε να δούμε πιο αναλυτικά για ποιο λόγο παρουσιάζεται αυτό το πρόβλημα. Η σχέση κανονικών εξισώσεων της μεθόδου αυτοσυσχέτισης ήταν η παρακάτω,

$$\sum_{k=1}^{p} a_k R_s[i-k] = R_s[i], \quad i = 1, 2.., p \tag{2.30}$$

Το μοντελοποιημένο σήμα που υπολογίζουμε $\hat{s}[n]$ προφανώς ικανοποιεί το μοντέλο μας

και την παραπάνω εξίσωση, με τους ίδιους συντελεστές a_k. Προκύπτει έτσι ότι πρέπει να ισχύει η παρακάτω σχέση:

$$R_s[i] = cR_{\hat{s}}[i], \quad i = 0, 1, .., p \tag{2.31}$$

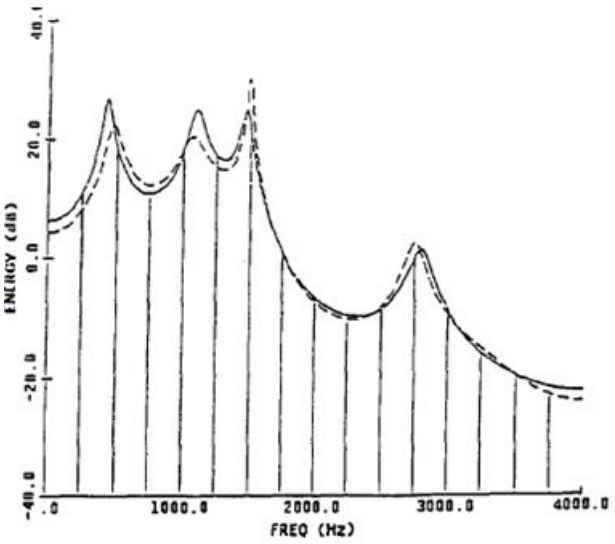
Στην περίπτωση ωστόσο των έμφωνων ήχων, το R_s αντιστοιχεί στην αυτοσυσχέτιση του τελικού σήματος φωνής του οποίου το φάσμα έχει προκύψει έπειτα από πολλαπλασιασμό στο πεδίο των συχνοτήτων με την παλμοσειρά της περιοδικής διέγερσης. Με τον πολλαπλασιασμό αυτό ουσιαστικά προκύπτει μια δειγματοληπτημένη μορφή του φάσματος του φίλτρου, και αυτό έχει σαν συνέπεια η αυτοσυσχέτιση του τελικού σήματος να ισούται με μια επικαλυπτόμενη εκδοχή της αρχικής αυτοσυσχέτισης.

Θα ισχύει δηλαδή ότι

$$R_s[i] = \sum_{l=-\infty}^{\infty} R_{org}[i - lT] \tag{2.32}$$

όπου T είναι η περίοδος της περιοδικής διέγερσης και R_{org} η αρχική αυτοσυσχέτιση.

Φαίνεται λοιπόν ότι η μέθοδος Autocorrelation προσπαθεί να ταιριάξει την αυτοσυσχέτιση του σήματος μοντελοποίησης $\hat{s}$, με μια παραμορφωμένη εκδοχή της αυτοσυσχέτισης του πραγματικού σήματος. Μάλιστα όσο είναι μεγαλύτερο το pitch του ήχου (δηλαδή μικρότερη η περίοδος), τόσο αυξάνεται η παραμόρφωση. Γί αυτό και παρουσιάζεται εντονότερο το πρόβλημα σε ήχους με μεγάλο pitch.



Σχήμα 2.13: Φάσμα συνάρτησης αυτοσυσχέτισης περιοδικού σήματος [34]

2.4.2 Μέθοδος μοντελοποίησης με διακριτό φίλτρο με μόνο πόλους (Discrete All-Pole Modeling)

Σύμφωνα με τους El-Jaroudi, J.Makhoul, η αδυναμία του μοντέλου αυτοσυσχέτισης είναι εγγενής με το κριτήριο σφάλματος το οποίο χρησιμοποιεί. Οι ίδιοι ανέπτυξαν μια νέα μέθοδο για τον υπολογισμό του φάσματος του φίλτρου της φωνητικής οδού την οποία ονόμασαν μέθοδο DAP (Discrete All Pole Modeling Method) [34].

Το νέο κριτήριο σφάλματος που όρισαν είναι το παρακάτω:

$$E_{IS} = \frac{1}{N}\sum_{m=1}^{N}(\frac{P(\omega_m)}{\hat{P}(\omega_m)} - ln\frac{P(\omega_m)}{\hat{P}(\omega_m)}) - 1 \tag{2.33}$$

όπου N είναι ο αριθμός των συχνοτήτων στις οποίες είναι υπολογισμένο το διακριτό φάσμα του σήματος φωνής το οποίο συμβολίζουμε με $P(w_m)$. Με $\hat{P}(w_m)$ συμβολίζουμε το φάσμα του μοντελοποιημένου σήματος.

Αποδεικνύεται ότι ελαχιστοποίηση του παραπάνω σφάλματος είναι ισοδύναμη με μεγιστοποίηση της ομαλότητας (flatness) του φάσματος λάθους $P(w_m)/\hat{P}(w_m)$, όπου η ομαλότητα ορίζεται ως το πηλίκο του γεωμετρικού μέσου των δειγμάτων προς τον αριθμητικό τους μέσο. Αυτό έχει σαν συνέπεια το φάσμα του λάθους να είναι όσο το δυνατόν πιο επίπεδο.

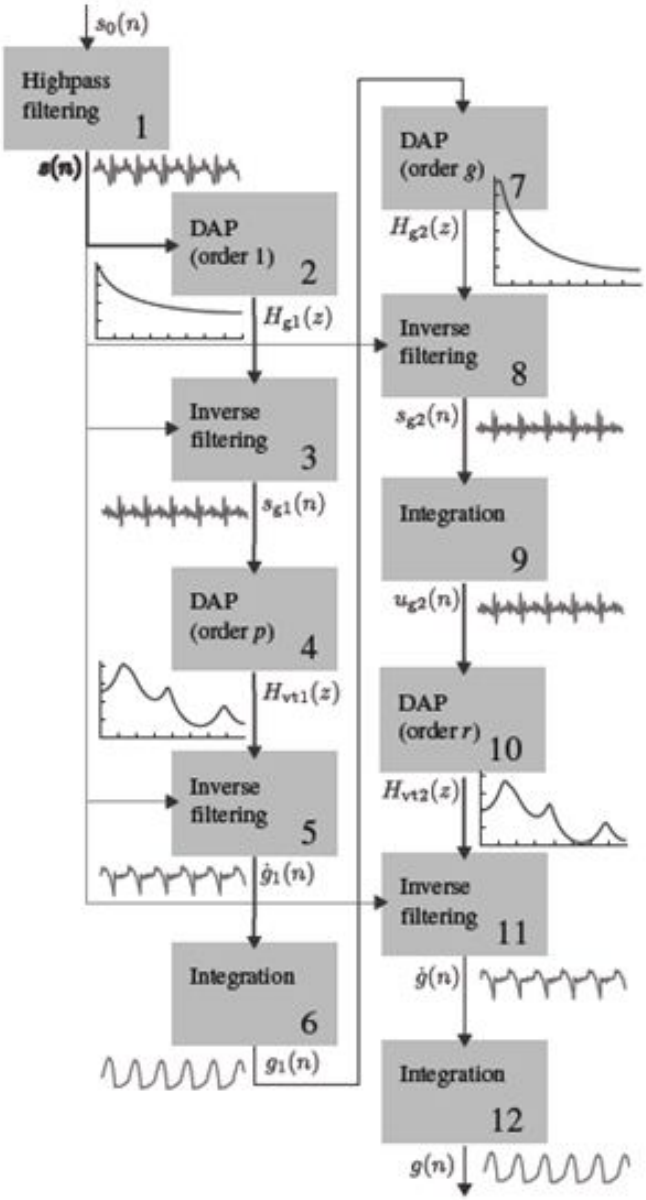
Απαιτώντας την ελαχιστοποίηση της συνάρτησης σφάλματος E_{IS}, προκύπτει όπως και πριν ότι πρέπει να ισχύει:

$$R_s[i] = cR_{\hat{s}}[i], \quad i = 0, 1, .., p \tag{2.34}$$

Η διαφορά με την μέθοδο γραμμικής πρόβλεψης, είναι ότι στην περίπτωση της γραμμικής πρόβλεψης, το $R_{\hat{s}}[i]$ ισούται με την αυτοσυσχέτιση του συνεχούς φάσματος του μοντέλου, ενώ στην περίπτωση της μεθόδου DAP που εξετάζουμε, το $R_{\hat{s}}[i]$ ισούται με την τιμή του δειγματοληπτημένου φάσματος του μοντέλου. Έτσι η μέθοδος DAP προσπαθεί να ταυτίσει την παραμορφωμένη από επικάλυψη αυτοσυσχέτιση του σήματος με την παραμορφτωμένη με τον ίδιο τρόπο από επικάλυψη αυτοσυσχέτιση του μοντέλου. Αυτή είναι και η ουσιαστική βελτίωση της μεθόδου αυτής σε σχέση με την μέθοδο αυτοσυσχέτισης, που την κάνει να δίνει καλύτερες εκτιμήσεις στην περίπτωση των έμφωνων ήχων. Λεπτομέρειες για τον τρόπο υπολογισμού των παραμέτρων a_i δίνονται στο [34]. Πρέπει επίσης να τονίσουμε ότι ο υπολογισμός των βέλτιστων συντελεστών που ελαχιστοποιούν το λάθος της μεθόδου DAP, έχει αυξημένο υπολογιστικό κόστος σε σχέση με τις προηγούμενες μεθόδους.

Αλγόριθμος IAIF για αντίστροφο φιλτράρισμα

Τα διάφορα στάδια της διαδικασίας αντίστροφου φιλτραρίσματος με την μέθοδο IAIF [35] φαίνονται συγκεντρωμένα στο σχήμα 2.15. Αρχικά στο 1ο στάδιο το σήμα φωνής περνάει μέσα από ένα υψιπερατό φίλτρο για να απαλείψει ενδεχόμενες παραμορφώσεις του μικροφώνου στις χαμηλές συχνότητες. Στα επόμενα στάδια, ο αλγόριθμος εφαρμόζει την διαδικασία του αντίστροφου φιλτραρίσματος επαναληπτικά σε 2 διαφορετικές φάσεις. Αρχικά εφαρμόζει τον αλγόριθμο DAP με τάξη φίλτρου ίση με 1, για να απαλείψει την επίδραση των χειλιών. Στη



Σχήμα 2.14: Σχηματικό διάγραμμα με τα διάφορα στάδια του IAIF αλγόριθμου [32]

συνέχεια αφού περάσει το σήμα από αυτό το φίλτρο, εφαρμόζει τον αλγόριθμο DAP με τάξη φίλτρου η οποία συνήθως είναι ίση με 2 φόρες τον ρυθμό δειγματοληψίας του σήματος σε kHz. Η έξοδος του φίλτρου αποτελεί την πρώτη εκτίμηση του σήματος πηγής, ενώ το φίλτρο που χρησιμοποιήθηκε την πρώτη εκτίμηση του φίλτρου. Στη συνέχεια ξεκινώντας από το νέο σήμα που προέκυψε, επαναλαμβάνει την ίδια διαδικασία, δηλαδή αφαιρεί πάλι την επίδραση των χειλιών και εφαρμόζει DAP ανάλυση για να αποκτήσει ακόμα καλύτερες προσεγγίσεις, κ.ο.κ.

2.4.3 Εξαγωγή χαρακτηριστικών από την κυματομορφή της πηγής

Από την στιγμή που ολοκληρώνεται η διαδικασία του αντίστροφου φιλτραρίσματος και διαθέτουμε την κυματομορφή της πηγής, μπορούμε να εξάγουμε διάφορα χαρακτηριστικά από αυτήν.

Χαρακτηριστικά στο πεδίο του χρόνου

Με την βοήθεια του εργαλείου TKK-Aparat, προσδιορίζουμε τις κρίσιμες χρονικές στιγμές στο σήμα της πηγής και στη συνέχεια με βάσει αυτές υπολογίζουμε διάφορα χαρακτηριστικά στο πεδίο του χρόνου. Όπως φαίνεται και σχήμα 2.16, οι κρίσιμες χρονικές στιγμές στο σήμα της πηγής είναι οι παρακάτω:

- T: Είναι η θεμελιώδης περίοδος του πλαισίου που εξετάζουμε και υπολογίζεται σύμφωνα με τον αλγόριθμο RAPT που αναλύσαμε. Οι διάφοροι παλμοί του πλαισίου απέχουν μεταξύ τους κατά T.

- t_{max}: Είναι η χρονική στιγμή που παρατηρείται η μέγιστη τιμή του πλαισίου. Η χρονική αυτή στιγμή αντιστοιχεί στην στιγμή μέγιστης ροής αέρα σε κάποιον από τους γλωττιδικούς παλμούς του πλαισίου.

- t_{o1}, t_{o2}: Επειδή οι φωνητικές χορδές ανοίγουν προοδευτικά, προσδιορίζουμε δύο χρονικές στιγμές που σχετίζονται με το άνοιγμα τους. Την πρώτη ή κύρια t_{o1}, και την δευτερεύουσα t_{o2}.

- t_c: Είναι η χρονική στιγμή που αντιστοιχεί στο κλείσιμο των φωνητικών χορδών.

- t_{min}: Είναι η χρονική στιγμή της ελάχιστης ροής αέρα από τις φωνητικές χορδές και εντοπίζεται λίγο μετά το t_c.

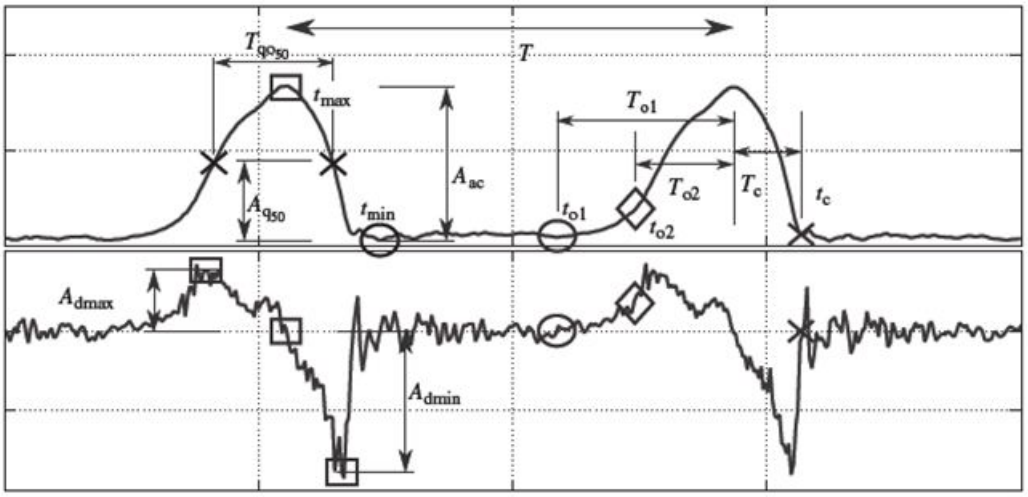
- t_{qo1}, t_{qo2}: Είναι οι χρονικές στιγμές εκατέρωθεν του t_{max} που αντιστοιχούν σε πλάτος σήματος ίσο με το μισό του μέγιστου.

Χρησιμοποιώντας τις παραπάνω κρίσιμες χρονικές στιγμές που αναφέραμε μαζί και με άλλες που φαίνονται στο σχήμα, υπολογίζουμε μια σειρά από χαρακτηριστικά. Παραθέτουμε παρακάτω μερικά από τα πιο βασικά.

- $OQ = (t_c - t_{o1})/T$: Λόγος ανοίγματος των φωνητικών χορδών (*Open Quotient*) . Μετράει το ποσοστό της ανοιχτής φάσης σε σχέση με τον συνολικό κύκλο της περιόδου.

- $SQ = (t_{max} - t_{o1})/(t_c - t_{max})$: Λόγος ταχύτητας (*Speed Quotient*) . Μετράει τον λόγο της διάρκειας της ανοιχτής φάσης προς την διάρκεια της κλειστής φάσης.

- $AQ = A_{ac}/A_{dmin}$: Λόγος πλάτους (*Amplitude Quotient*). Μετράει τον λόγο του μέγιστου πλάτους του σήματος του γλωττιδικού παλμού προς την ελάχιστη τιμή της παραγώγου του.



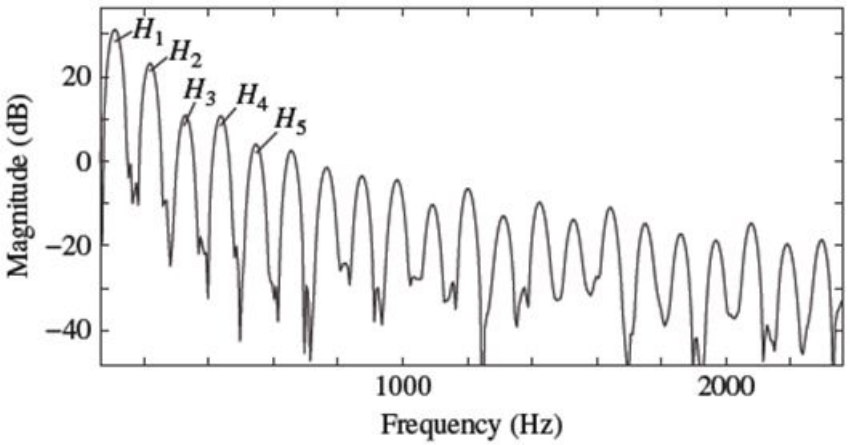
Σχήμα 2.15: Πλαίσιο ανάλυσης και μέτρηση παραμέτρων του γλωττιδικού παλμού [32]

- $NAQ = AQ/T$: Κανονικοποιημένος λόγος πλάτους με βάση την περίοδο.

Χαρακτηριστικά στο πεδίο συχνοτήτων

Για να εξάγουμε χαρακτηριστικά στο πεδίο των συχνοτήτων υπολογίζουμε αρχικά το φάσμα του γλωττιδικού παλμού που έχουμε στην διάθεση μας. Από το φάσμα του σήματος της πηγής εξάγουμε δύο βασικά χαρακτηριστικά:

- $H_1 - H_2$: Είναι απλά η διαφορά των πλατών της θεμελιώδους συχνότητας (*pitch*) H_1 και της πρώτης αρμονικής της H_2, μετρημένων σε *decibels* όπως φαίνεται στο παρακάτω σχήμα.



Σχήμα 2.16: Αρμονικές του φάσματος γλωττιδικού παλμού [32]

- $HRF = \sum_{k\geq 2} H_k/H_1$: Παράγοντας "ποσότητας" αρμονικού περιεχομένου (Harmonic Richness Factor). Μετράει τον λόγο του αθροίσματος των πλατών των αρμονικών προς το πλάτος της θεμελιώδους συχνότητας.

Επισημαίνουμε τέλος ότι για η εφαρμογή του αλγορίθμου IAIF προϋποθέτει την ύπαρξη ενός αλγορίθμου που θα αποφασίζει αν το εκάστοτε τμήμα φωνής είναι έμφωνο και θα προσδιορίζει την θεμελιώδη συχνότητα του. Στην περίπτωση μας χρησιμοποιούμε τον αλγόριθμο RAPT.

Αφού λοιπόν εντοπίσουμε τα έμφωνα τμήματα του σήματος, υπολογίζουμε τον γλωττιδικό παλμό που τους αντιστοιχεί και εξάγουμε από αυτόν χαρακτηριστικά στο πεδίο του χρόνου και των συχνοτήτων, έτσι ώστε να έχουμε μια καλή περιγραφή της μορφής του.

2.5 Χαρακτηριστικά AM-FM

2.5.1 Μη γραμμικά φαινόμενα στην παραγωγή φωνής

Όπως αναφέραμε προηγουμένως το μοντέλο πηγής-φίλτρου κάνοντας κάποιες απλοποιητικές παραδοχές καταφέρνει να προσεγγίσει τον μηχανισμό παραγωγής της φωνής και να παράγει ικανοποιητικά αποτελέσματα σε αρκετές περιπτώσεις. Μάλιστα ο αλγόριθμος εξαγωγής των χαρακτηριστικών του γλωττιδικού παλμού που παρουσιάσαμε στην προηγούμενη ενότητα στηρίζεται σε αυτό το μοντέλο. Ωστόσο το γραμμικό μοντέλο πηγής-φίλτρου δεν παύει να αποτελεί μια προσέγγιση του πραγματικού μηχανισμού παραγωγής της φωνής. Έρευνες έχουν δείξει ότι η παραγωγή της φωνής αποτελεί ένα περίπλοκο μηχανισμό στον οποίο λαμβάνουν χώρα μια σειρά από μη-γραμμικά και ταχύτατα χρονικά μεταβαλλόμενα φαινόμενα τα οποία αντιτίθενται στις παραδοχές του προηγούμενου μοντέλου [36, 37, 38]. Πιο συγκεκριμένα παραθέτουμε τις παρακάτω ενδείξεις.

- Η φωνητική οδός και η πηγή δεν είναι ανεξάρτητες αλλά αλληλεπιδρούν κατά την διάρκεια της ομιλίας, και κάποιες φορές σε μη αμελητέο βαθμό.
- Το κύμα του αέρα της πηγής δεν ρέει ομοιόμορφα κατά μήκος μιας διατομής της φωνητικής οδού σαν ένα μονοδιάστατο σήμα, ενώ μάλιστα αρκετές φορές αποσπώνται από αυτό κομμάτια τα οποία δημιουργούν μικρούς στρόβιλους που επηρεάζουν την ροή του.
- Ο σχηματισμός της φωνητικής οδού και επομένως οι συντονισμοί της μεταβάλλονται αρκετά γρήγορα με αποτέλεσμα να μην μπορούν να θεωρηθούν σταθεροί ούτε κατά την διάρκεια μιας περιόδου του γλωττιδικού παλμού.

Όλες αυτές οι ενδείξεις (και άλλες ακόμα), θέτουν ένα άνω όριο στην απόδοση και την ακρίβεια του μοντέλου πηγής-φίλτρου. Δημιουργήθηκε η ανάγκη λοιπόν να υπάρξει μια μέτρηση αυτών των μη-γραμμικών φαινομένων τα οποία παίζουν σημαντικό ρόλο στην διαμόρφωση της ποιότητας της φωνής. Έτσι οι Maragos, Quatieri, Kaiser [39, 40, 41, 42], πρότειναν ένα νέο, μη γραμμικό μοντέλο για την φωνή το οποίο ονομάζεται μοντέλο διαμόρφωσης AM-FM.

2.5.2 Μοντέλο AM-FM

Σύμφωνα με το μοντέλο AM-FM, κάθε συντονισμός (formant) ενός σήματος φωνής αναπαρίσταται σαν ένα σήμα AM-FM, δηλαδή σαν ένα σήμα διαμορφωμένο ταυτόχρονα κατά πλάτος και κατά συχνότητα. Αν $r(t)$ είναι ένας συντονισμός έχουμε:

$$r(t) = a(t)\cos\underbrace{(2\pi(\omega_c t + \omega_m \int_0^t q(\tau)d\tau + \phi(0)))}_{\phi(t)}, \tag{2.35}$$

όπου $\mathrm{f}_c = \frac{\omega_c}{2\pi}$, είναι η κεντρική συχνότητα του συντονισμού. Όπως φαίνεται από την παραπάνω εξίσωση η συχνότητα του συντονισμού δεν είναι σταθερή για το δεδομένο πλαίσιο ανάλυσης, αλλά μεταβάλλεται γύρω από μια κεντρική συχνότητα. Παράλληλα μεταβάλλεται και το πλάτος

της $a(t)$ σαν συνάρτηση του χρόνου. Η στιγμιαία συχνότητα του συντονισμού ορίζεται ως η χρονική παράγωγος της φάσης $\phi(t)$:

$$f(t) = \frac{1}{2\pi}[f_c + f_m q(t)], \quad q(t) \in [-1, 1], \tag{2.36}$$

όπου f_m είναι η μέγιστη απόκλιση της στιγμιαίας συχνότητας από την κεντρική συχνότητα. Μπορεί τώρα κάποιος να αναπαραστήσει το σήμα φωνής ως την επαλληλία από N διαφορετικά σήματα συντονισμών, οπότε έχουμε:

$$s(t) = \sum_{k=1}^{N} r_k(t), \tag{2.37}$$

όπου N είναι ο αριθμός των formants της φωνητικής οδού για το συγκεκριμένο τμήμα φωνής.

Το μοντέλο AM-FM αναπαριστώντας το σήμα φωνής με αυτόν τον τρόπο, αποφεύγει να κάνει τον διαχωρισμό πηγής-φίλτρου ή να υποθέσει οποιαδήποτε ανεξαρτησία μεταξύ τους. Αποκτά έτσι μεγαλύτερη ελευθερία για να μελετήσει και να μετρήσει τις συνέπειες των μη γραμμικών φαινομένων στο σήμα φωνής. Στη συνέχεια παρουσιάζουμε τα βήματα του αλγορίθμου για τον υπολογισμό των συνιστωσών $a(t)$ και $f(t)$ των AM-FM σημάτων συντονισμού.

2.5.3 Σχήμα αποδιαμόρφωσης AM-FM

Ο αλγόριθμος υπολογισμού των συνιστωσών $a(t)$ και $f(t)$ για το κάθε σήμα συντονισμού αποτελείται από δύο βασικά βήματα. Αρχικά κατασκευάζουμε μια συστοιχία από ζωνοπερατά φίλτρα Gabor τα κέντρα των οποίων ακολουθούν την κλίμακα Mel που αναφέραμε και προηγουμένως.

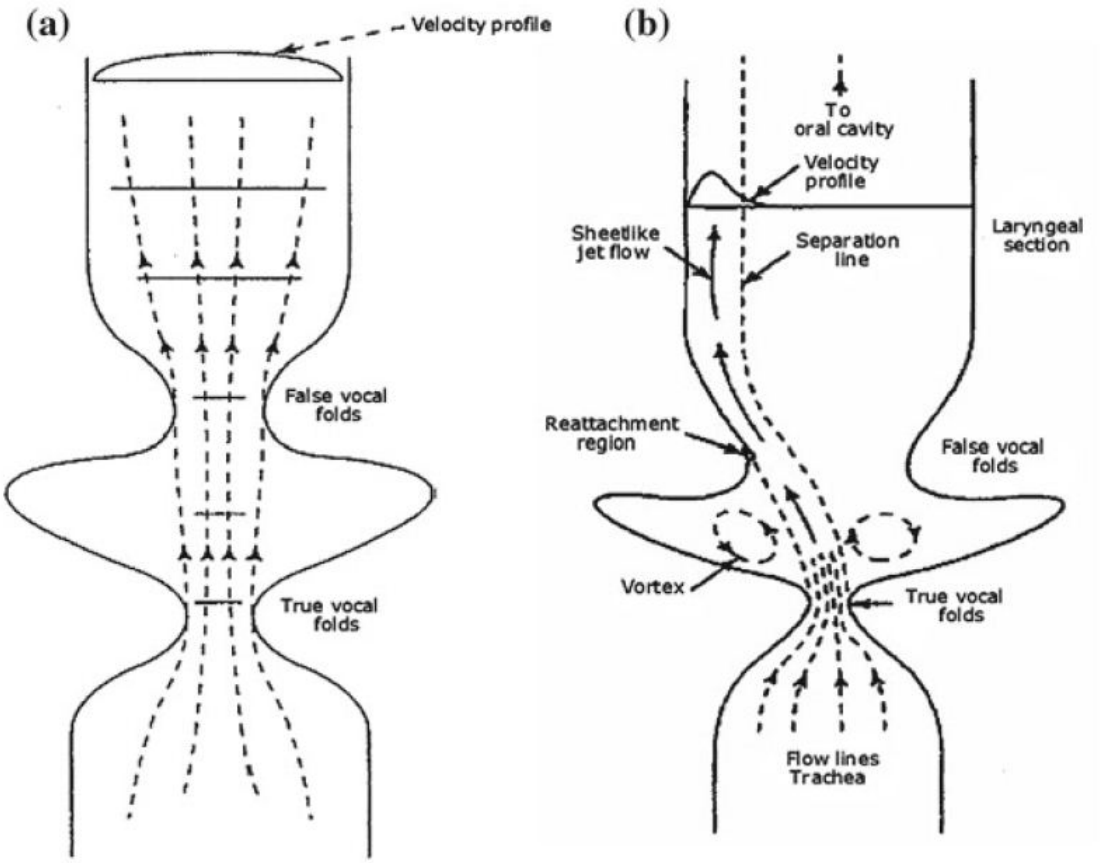
Για τα πραγματικά φίλτρα Gabor που χρησιμοποιούμε, η κρουστική τους απόκριση στο πεδίο του χρόνου και των συχνοτήτων δίνονται από τους παρακάτω τύπους:

$$h(t) = e^{-a^2 t^2} \cos(2\pi v t) \tag{2.38}$$

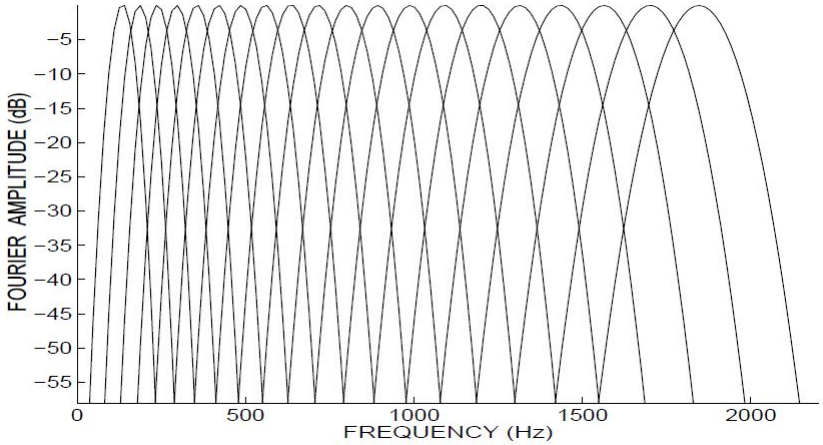
$$H(w) = \frac{\sqrt{\pi}}{2a}(e^{[-\frac{\pi^2(w-v)^2}{a^2}]} + e^{[-\frac{\pi^2(w+v)^2}{a^2}]}) \tag{2.39}$$

όπου v είναι η κεντρική συχνότητα του φίλτρου η οποία ισούται με την κεντρική συχνότητα του συντονισμού, και α είναι η παράμετρος που ρυθμίζει το εύρος του φίλτρου. Τα φίλτρα Gabor επιλέχθηκαν εξαιτίας της βέλτιστης διακριτικής τους ικανότητας τόσο στο πεδίο του χρόνου όσο και της συχνότητας. Στο πείραμα μας ακολουθούμε την μέθοδο των Maragos,Dimitriadis [45], και χρησιμοποιούμε συστοιχία από συνολικά 6 φίλτρα Gabor τα οποία έχουν επικάλυψη 50% μεταξύ τους.

Η εφαρμογή καθενός από τα φίλτρα Gabor της συστοιχίας στο σήμα μας, δίνει σαν αποτέλεσμα ένα σήμα συντονισμού $r_i(t)$. Μετά την απόκτηση των σημάτων συντονισμού, το



Σχήμα 2.17: (α)Γραμμικό και (β) μη-γραμμικό μοντέλο παραγωγής φωνής [43]



Σχήμα 2.18: Συστοιχία φίλτρων Gabor [44]

δεύτερο βήμα είναι ο υπολογισμός του στιγμιαίου πλάτους και συχνότητας $a(t)$ και $f(t)$ σε καθένα από αυτά. Το βήμα αυτό ονομάζεται αποδιαμόρφωση και επιτυγχάνεται με την χρήση του αλγορίθμου ESA (energy separator algorithm). Ο αλγόριθμος ESA στηρίζεται στην χρήση του τελεστή ενέργειας Teager ο οποίος ορίζεται ως εξής:

$$\Psi_c[x(t)] = [\dot{x}(t)]^2 - x(t)\ddot{x}(t) \tag{2.40}$$

Ο τελεστής Teager μετράει την ενέργεια ενός γραμμικού ταλαντωτή μάζας m και σταθεράς ταλάντωσης k του οποίου η διαφορική εξίσωση κίνησης δίνεται από την σχέση: $\ddot{x}(t) + (k/m)x(t) = 0$. Η γενική λύση της εξίσωσης αυτής είναι $x(t) = A\cos(\omega_o t + \theta)$, όπου $\omega_o = \sqrt{k/m}$, ενώ η στιγμιαία ενέργεια E_o του ταλαντωτή είναι σταθερή και ίση με το άθροισμα κινητικής και δυναμικής ενέργειας και ισούται με:

$$E_o = \frac{m}{2}(\dot{x})^2 + \frac{k}{2}x^2 = \frac{m}{2}(A\omega_o)^2. \tag{2.41}$$

Έτσι αν εφαρμόσουμε τον τελεστή ενέργειας Teager θα πάρουμε:

$$\Psi_c[x(t)] = A^2\omega_o^2 = \frac{E_o}{m/2}, \tag{2.42}$$

δηλαδή την ενέργεια ανά μονάδα μάζας. Ομοίως αν εφαρμόσουμε τον τελεστή Teager σε ένα AM-FM σήμα θα έχουμε:

$$\Psi_c[a(t)cos(\phi(t))] = \overbrace{(a\dot{\phi})^2}^{D} + \overbrace{\frac{\Psi_c(a)}{2}}^{E_L} + \overbrace{\frac{a^2\ddot{\phi}}{2}sin(2\phi) + \frac{\Psi_c(a)}{2}cos(2\phi)}^{E_H}. \tag{2.43}$$

Το αποτέλεσμα αυτό μπορεί να θεωρηθεί προσεγγιστικά ίσο με $(a\dot{\phi})^2$ όπως και προηγουμένως, εφόσον οι όροι E_L και E_H είναι μικροί σε σχέση με το D. Για να συμβεί αυτό εφαρμόζουμε βαθυπερατό φιλτράρισμα στην έξοδο του τελεστή Teager, ώστε να απομακρύνουμε τον κυριότερο ανεπιθύμητο όρο E_H ο οποίος περιέχει μια συχνότητα διπλάσια από την κεντρική f_c. Μπορούμε τότε να γράψουμε ότι:

$$\Psi_c[r_i(t)] \approx \frac{1}{2\pi}(a(t)f(t))^2 \tag{2.44}$$

Σύμφωνα με τον αλγόριθμο ESA, για τον διαχωρισμό του γινομένου ενέργειας σε $f(t)$ και $a(t)$ χρησιμοποιούνται οι δύο παρακάτω σχέσεις:

$$f(t) = \frac{1}{2\pi}\sqrt{\frac{\Psi(\dot{x}(t))}{\Psi(x(t))}} \tag{2.45}$$

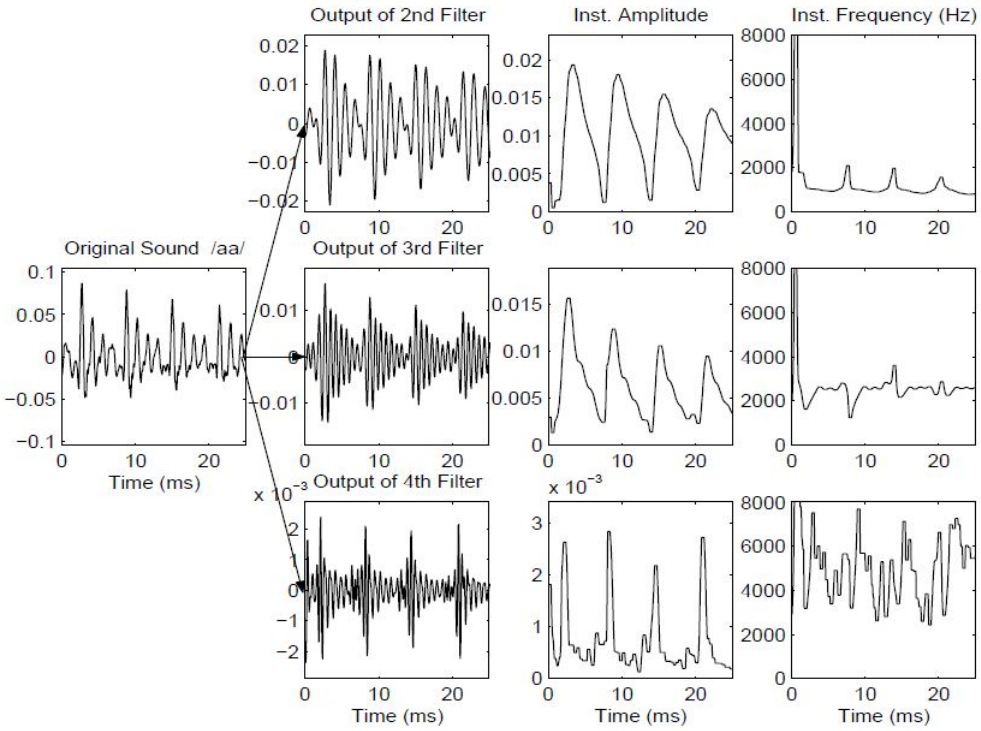
$$|a(t)| = \frac{\Psi(x(t))}{\sqrt{\Psi(\dot{x}(t))}} \tag{2.46}$$

Η προηγούμενη ανάλυση αναφέρεται σε σήματα συνεχούς χρόνου. Υπάρχει ανάλογη διαδικασία για σήματα διακριτού χρόνου στα οποία η παράγωγος υπολογίζεται ως $x''[n] = x[n]-x[n-1]$, οπότε προκύπτουν και αντίστοιχοι τύποι για τον υπολογισμό των $f[n]$ και $|a[n]|$.

Στην δική μας έρευνα ωστόσο χρησιμοποιούμε τον αλγόριθμο Gabor ESA [45] ο οποίος επεκτείνει αρχικά το σήμα διακριτού χρόνου στο συνεχές πεδίο και στη συνέχεια εφαρμόζει εκεί τον αλγόριθμο ESA. Με αυτό τον τρόπο αντί για τον θορυβώδη υπολογισμό της παραγώγου από την εξίσωση διαφορών στον διακριτό χρόνο, πραγματοποιείται μια πιο ομαλή και ακριβής εκτίμηση της παραγώγου του σήματος. Χρησιμοποιώντας μάλιστα την προσεταιριστική ιδιότητα της συνέλιξης ως προς την χρονική παράγωγο, έχουμε:

$$\Psi[x(t) * g(t)] = [x(t) * \frac{dg(t)}{dt}]^2 - (x(t) * g(t))[x(t) * \frac{d^2g(t)}{dt^2}] \tag{2.47}$$

Οπότε απλά πραγματοποιούμε συνέλιξη του σήματος με την χρονική παράγωγο του φίλτρου Gabor (η οποία είναι γνωστή) και πετυχαίνουμε πιο ακριβή και ομαλή εκτίμηση του τελεστή Teager. Παρακάτω δείχνουμε σχηματικά την AM-FM ανάλυση και τα $a(t)$ και $f(t)$ σήματα που προκύπτουν στους διάφορους συντονισμούς ενός σήματος φωνής.



Σχήμα 2.19: Ανάλυση σήματος σε AM-FM συνιστώσες [46]

2.5.4 Χαρακτηριστικά AM-FM

Έχοντας υπολογίσει τις συνιστώσες του στιγμιαίου πλάτους $|a(t)|$ και της στιγμιαίας συχνότητας $f(t)$ για καθένα από τα 6 σήματα συντονισμού $\mathrm{r}_i(t)$, τις χρησιμοποιούμε για να εξάγουμε τα παρακάτω χαρακτηριστικά:

- Σταθμισμένη μέση στιγμιαία συχνότητα F_{wi} που δίνεται από την σχέση:

$$F_{wi} = \frac{\int_0^T f_i(t) a_i^2(t) dt}{\int_0^T a_i^2(t) dt} \tag{2.48}$$

  Το F_w είναι η μέση συχνότητα στο χρονικό διάστημα του πλαισίου που εξετάζουμε, σταθμισμένη με το τετράγωνο του πλάτους $|a(t)|$.

- Σταθμισμένο εύρος στιγμιαίας συχνότητας B_i που δίνεται από την σχέση:

$$B_i = \frac{\int_0^T [\dot{a}_i^2(t) + (f_i(t) - F_{wi})^2 a_i^2(t)] dt}{\int_0^T a_i^2(t) dt} \tag{2.49}$$

- Ποσοστό διαμόρφωσης συχνότητας FMP_i, ίσο με τον λόγο των προηγούμενων δύο όρων.

$$FMP_i = \frac{B_i}{F_{wi}} \tag{2.50}$$

- Μέση τιμή του στιγμιαίου πλάτους $|a(t)|$.

  Η μεταβλητή αυτή μας δίνει μια εικόνα για την μέση τιμή του πλάτους του συντονισμού λαμβάνοντας υπόψιν μη γραμμικά φαινόμενα όπως οι παλμοί διαμόρφωσης που υπάρχουν μέσα σε μια περίοδο pitch (Σχήμα 2.19).

Κεφάλαιο 3

Μέθοδοι επιλογής χαρακτηριστικών

Η επιλογή χαρακτηριστικών αποτελεί ένα πολύ κρίσιμο βήμα στα προβλήματα αναγνώρισης προτύπων. Στόχος της είναι, δεδομένου του αρχικού διανύσματος χαρακτηριστικών, να επιλέξει τα χαρακτηριστικά εκείνα τα οποία σχετίζονται περισσότερο με το πρόβλημα. Με τον τρόπο αυτό αυξάνεται η απόδοση του συστήματος και αντιμετωπίζεται το πρόβλημα της διαστασιμότητας. Η επιλογή χαρακτηριστικών έχει παρόμοιο ρόλο με τις μεθόδους μετασχηματισμού τους (π.χ. PCA, LDA), με την διαφορά ότι επιλέγεται απλά ένα υποσύνολο από τα ήδη υπάρχοντα χαρακτηριστικά. Το πλεονέκτημα αυτής της μεθόδου είναι ότι διατηρείται η φυσική σημασία των αρχικών χαρακτηριστικών και διερευνάται η σχετικότητα τους με το πρόβλημα.

Για την επιλογή των χαρακτηριστικών χρειάζεται μια στρατηγική αναζήτησης υποσυνόλων καθώς και μια αντικειμενική συνάρτηση για την αξιολόγηση τους. Ανάλογα με την συνάρτηση-κριτήριο, με βάση το οποίο γίνεται η επιλογή, προκύπτουν δύο βασικές κατηγορίες αλγορίθμων. Οι αλγόριθμοι επιλογής με βάση το ποσοστό αναγνώρισης του ταξινομητή και οι αλγόριθμοι επιλογής με βάση την εφαρμογή κάποιου ειδικού φίλτρου. Οι αλγόριθμοι φίλτρου έχουν καλύτερες ιδιότητες γενίκευσης καθώς δεν εξαρτώνται από το σύστημα αναγνώρισης που χρησιμοποιούμε. Από την άλλη, οι αλγόριθμοι που στηρίζονται στο ποσοστό αναγνώρισης μπορούν να επιλέξουν χαρακτηριστικά τα οποία συνδυάζονται καλύτερα με το συγκεκριμένο σύστημα που διαθέτουμε.

3.1 Αλγόριθμοι επιλογής με βάση το ποσοστό αναγνώρισης (wrappers)

Οι αλγόριθμοι αυτοί αξιολογούν ένα υποσύνολο χαρακτηριστικών με βάση το ποσοστό αναγνώρισης που πετυχαίνει το σύστημα ταξινόμησης χρησιμοποιώντας το. Οι δύο πιο κλασσικοί και κυριότεροι αλγόριθμοι αυτής της κατηγορίας είναι η προς τα εμπρός σειριακή επιλογή (FSS-forward sequential selection) και η προς τα πίσω σειριακή επιλογή (BSS-backward sequential selection).

3.1.1 Επιλογή προς τα εμπρός (Forward Selection)

Έστω το αρχικό σύνολο χαρακτηριστικών S_n μεγέθους n. Ο αλγόριθμος forward selection επιλέγει m χαρακτηριστικά από το αρχικό σύνολο με έναν αυξητικό και ακολουθητικό τρόπο. Συγκεκριμένα σε κάθε επανάληψη προσθέτει στο σύνολο το χαρακτηριστικό το οποίο μεγιστοποιεί το ποσοστό αναγνώρισης του προβλήματος, αν συνδυαστεί με τα υπόλοιπα χαρακτηριστικά που ήδη έχουν επιλεχθεί. Αρχικά το σύνολο είναι κενό και το πρώτο χαρακτηριστικό που επιλέγεται είναι αυτό που από μόνο του μεγιστοποιεί το ποσοστό αναγνώρισης. Ο αλγόριθμος σταματάει μόλις φθάσει στα m χαρακτηριστικά. Παρακάτω φαίνονται τα βήματα σε μορφή ψευδοκώδικα:

1. Αρχικοποίηση με το κενό σύνολο $Y_0 = \{\varnothing\}$, k=0
2. Επιλογή του επόμενου χαρακτηριστικού $x^+ = argmax_{x \notin Y_k} Sucess\_Rate[Y_k \cup \{x\}]$
3. Ανανέωση συνόλου $Y_{k+1} = Y_k \cup \{x^+\}$, k=k+1
4. Αν $k < m$ επιστροφή στο βήμα 2

Σημειώσεις:

- Ο αλγόριθμος FSS λειτουργεί καλύτερα όταν το μέγεθος m του ζητούμενου υποσυνόλου είναι μικρό καθώς αρχικά το εύρος αναζήτησης του είναι μεγαλύτερο.
- Το κύριο μειονέκτημα της μεθόδου αυτής είναι ότι δεν είναι δυνατή η επαναξιολόγηση και απομάκρυνση ενός χαρακτηριστικού αφότου αυτό έχει προστεθεί στο σύνολο. Συχνά π.χ. κάποια χαρακτηριστικά καθίστανται περιττά έπειτα από την πρόσθεση νέων.

Πρέπει να αναφέρουμε ότι ο αλγόριθμος FSS αποτελεί μια στρατηγική αναζήτησης στην οποία εκτός από το τελικό ποσοστό αναγνώρισης μπορεί να εφαρμοστεί και οποιαδήποτε άλλη αντικειμενική συνάρτηση (πχ. συσχέτιση μεταξύ των χαρακτηριστικών). Ωστόσο αποτελεί έναν από τους πιο κλασσικούς αλγορίθμους τύπου wrapper, για αυτό και τον αναφέρουμε σε αυτή την κατηγορία.

3.1.2 Επιλογή προς τα πίσω (Backward Selection)

Ο αλγόριθμος αυτός είναι όμοιος με τον προηγούμενο αλλά ακολουθεί αντίστροφη διαδικασία. Ξεκινώντας από το αρχικό σύνολο χαρακτηριστικών που διαθέτουμε, αφαιρεί κάθε φορά ένα χαρακτηριστικό έτσι ώστε να μεγιστοποιείται το ποσοστό αναγνώρισης στο σύνολο που απομένει:

1. Αρχικοποίηση με το αρχικό σύνολο $Y_0 = \{S_n\}$, k=n
2. Απομάκρυνση του επόμενου χαρακτηριστικού $x^- = argmax_{x \in Y_k} Sucess\_Rate[Y_k - \{x\}]$
3. Ανανέωση συνόλου $Y_{k+1} = Y_k - \{x^-\}$, k=k-1
4. Αν $k > m$ επιστροφή στο βήμα 2

Πρέπει να σημειωθεί ότι:

- Ο αλγόριθμος BSS λειτουργεί καλύτερα όταν το μέγεθος m του ζητούμενου υποσυνόλου είναι μεγάλο καθώς αρχικά το εύρος αναζήτησης του είναι μεγαλύτερο.

- Το κύριο μειονέκτημα της μεθόδου αυτής είναι ότι δεν είναι δυνατή η επαναξιολόγηση ενός χαρακτηριστικού αφότου αυτό έχει απομακρυνθεί.

Πρέπει να πούμε ότι οι μέθοδοι FSS και BSS αποτελούν κατά βάση άπληστους αλγορίθμους και συχνά εγκλωβίζονται σε τοπικά μέγιστα του ποσοστού επιτυχίας για το δεδομένο μέγεθος υποσυνόλου m.

3.1.3 Επιλογή με προσθήκη και αφαίρεση χαρακτηριστικών σε κάθε βήμα

Η μέθοδος αυτή προσπαθεί να καλύψει τις αδυναμίες των προηγούμενων δύο (FSS, BSS), προσφέρωντας την ικανότητα επαναξιολόγησης χαρακτηριστικών που προστέθηκαν ή απομακρύνθηκαν από το σύνολο προηγουμένως. Συγκεκριμένα σε κάθε επανάληψη προσθέτει έναν αριθμό από L χαρακτηριστικά και στη συνέχεια αφαιρεί R.

1. Αν $L > R$ τότε

Αρχικοποίηση με το κενό σύνολο $\mathrm{Y}_0 = \{\varnothing\}, \quad k = 0$

αλλιώς

Αρχικοποίηση με το αρχικό σύνολο $\mathrm{Y}_0 = \{S_n\}$

2. Επανάληψη L φορές $\mathrm{x}^+ = argmax_{x \notin Y_k} Sucess\_Rate[Y_k \cup \{x\}]$

$\mathrm{Y}_{k+1} = Y_k \cup \{x^+\}, \quad k = k + 1$

3. Επανάληψη R φορές $\mathrm{x}^- = argmax_{x \in Y_k} Sucess\_Rate[Y_k - \{x\}]$

$\mathrm{Y}_{k+1} = Y_k - \{x^-\}, \quad k = k - 1$

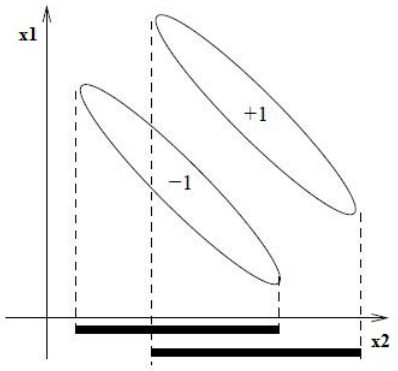
4. Αν $k < m$ επιστροφή στο βήμα 2

Σημειώσεις:

- Το κύριο μειονέκτημα αυτής της μεθόδου είναι ότι υπάρχει έλλειψη γνώσης για τους κατάλληλους αριθμούς L και R που πρέπει να χρησιμοποιηθούν ώστε να υπάρξει καλύτερο αποτέλεσμα.

3.2 Αλγόριθμοι επιλογής με εφαρμογή ειδικού φίλτρου (filters)

Οι αλγόριθμοι αυτής της κατηγορίας χρησιμοποιούν ως κριτήριο για την αξιολόγηση των υποσυνόλων συναρτήσεις ανεξάρτητες με το ποσοστό αναγνώρισης. Συνήθως οι συναρτήσεις αυτές μετράνε μεγέθη σχετικά με την θεωρία πληροφορίας, όπως είναι η αμοιβαία πληροφορία μεταξύ των χαρατηριστικών ή η πληροφορία που δίνουν τα χαρακτηριστικά για το πρόβλημα. Κύριος σκοπός των μεθόδων αυτών είναι να επιλεγούν χαρακτηριστικά τα οποία να είναι όσο το δυνατόν περισσότερο ασυσχέτιστα μεταξύ τους, και παράλληλα, σχετικά με το πρόβλημα.



Σχήμα 3.1: Παράδειγμα χαρακτηριστικών με χαμηλά F-scores [47]

3.2.1 Αλγόριθμος επιλογής με βάση το F-score

Ο αλγόριθμος F-score αποτελεί μια απλή μέθοδο για την μέτρηση της διαφορετικότητας δύο συνόλων πραγματικών αριθμών. Δεδομένων των διανυσμάτων χαρακτηριστικών των δειγμάτων εκπαίδευσης x_k, $k = 1, ..., m$, έαν n_+ και n_- είναι ο αριθμός των δειγμάτων των δύο διαφορετικών κλάσεων τότε το F-score για το i-οστό χαρακτηριστικό ορίζεται ως:

$$F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \tag{3.1}$$

όπου τα $\bar{x}_i$, $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$ είναι αντίστοιχα οι μέσες τιμές του i-οστού χαρακτηριστικού συνολικά στα δείγματα, στα θετικά μόνο δείγματα και στα αρνητικά δείγματα. Με $\mathrm{x}_{k,i}^{(+)}$ συμβολίζουμε το i-οστό χαρακτηριστικό του k-οστού θετικού δείγματος (αντίστοιχα για τα αρνητικά). Ο αριθμητής είναι ένα μέτρο για το πόσο διαφέρουν τα θετικά από τα αρνητικά δείγματα, ενώ ο παρανομαστής μετράει την διαφορετικότητα στο εσωτερικό του κάθε συνόλου.

Όπως λογικά προκύπτει, όσο μεγαλύτερο είναι το F-score ενός χαρακτηριστικού, τόσο περισσότερο κατάλληλο κρίνεται για να διαχωρίσει τις δύο αυτές κλάσεις.

Ένα μειονέκτημα αυτής της μεθόδου είναι ότι δεν λαμβάνει υπόψιν την συμπληρωματική πληροφορία μεταξύ των χαρακτηριστικών που επιλέγει. Κάθε χαρακτηριστικό αξιολογείται ξεχωριστά, ανεξάρτητα από την πιθανή συνδιαστική δύναμη του με τα υπόλοιπα (Σχήμα 3.1).

Στο σχήμα 3.1, και τα δύο χαρακτηριστικά x_1, x_2 έχουν χαμηλά F-scores καθότι ο παρανομαστής (το άθροισμα των διασπορών των δύο κλάσεων) είναι αρκετά μεγαλύτερος από τον αριθμητή. Ωστόσο σε συνδιασμό τα δύο χαρακτηριστικά διαχωρίζουν ικανοποιητικά τις δύο κλάσεις.

Ο αλγόριθμος F-score αποτελεί μια πολύ απλή και γρήγορη μέθοδο που δίνει συχνά ικανοποιητικά αποτελέσματα. Συνήθως επιλέγεται ένα σύνολο χαρακτηριστικών των οποίων το F-score είναι μεγαλύτερο από κάποιο κατώφλι που ορίζουμε.

3.2.2 Αλγόριθμος μέγιστης σχετικότητας και ελάχιστου πλεονασμού (MRMR)

Ο αλγόριθμος MRMR προτάθηκε από τους Peng et. al [48] και στοχεύει στην επιλογή ενός υποσυνόλου του οποίου τα χαρακτηριστικά έχουν μέγιστη συσχέτιση με το πρόβλημα αναγνώρισης που εξετάζουμε, και ταυτόχρονα, ελάχιστη αμοιβαία πληροφορία μεταξύ τους. Συγκεκριμένα τα αρχικά του αλγορίθμου υποδηλώνουν δύο επιδιώξεις, την μέγιστη "σχετικότητα" και τον ελάχιστο "πλεονασμό". Στην συνέχεια παρουσιάζουμε με περισσότερη λεπτομέρεια τα βασικά σημεία του αλγορίθμου.

Ορίζουμε αρχικά την συνάρτηση αμοιβαίας πληροφορίας (mutual information) δύο διακριτών μεταβλητών ως:

$$I(x,y) = \sum_{i,j} p(x_i, y_j) log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \tag{3.2}$$

Χρησιμοποιούμε την συνάρτηση αμοιβαίας πληροφορίας ως ένα μέτρο ομοιότητας μεταξύ των δύο μεταβλητών. Η ιδέα πίσω από το κριτήριο του ελάχιστου ΄πλεονασμού΄ είναι να διαλέξουμε χαρακτηριστικά τα οποία έχουν μεταξύ τους μικρή αμοιβαία πληροφορία.

Αν με S συμβολίσουμε το σύνολο των επιλεγμένων χαρακτηριστικών από το αρχικό σύνολο Ω, τότε το μέτρο του πλεονασμού για αυτό το σύνολο δίνεται από τον τύπο:

$$W_S = \frac{1}{|S|^2} \sum_{i,j \in S} I(i,j) \tag{3.3}$$

όπου με $|S|$ συμβολίζουμε το πλήθος των χαρακτηριστικών του S. Το κριτήριο του ελάχιστου πλεονασμού αναζητά το υποσύνολο S το οποίο ελαχιστοποιεί το W_S.

Έστω ότι με C συμβολίζουμε το σύνολο των κλάσεων που εμφανίζονται στο πρόβλημα μας. Χρησιμοποιώντας και πάλι την συνάρτηση αμοιβαίας πληροφορίας, μπορούμε να μετρήσουμε την σχέση των χαρακτηριστικών ενός συνόλου με την μεταβλητή του συνόλου των κλάσεων C= $\{c_1, c_2, ...c_n\}$ του προβλήματος μας:

$$V_{C,S} = \frac{1}{|S|} \sum_{i \in S} I(C,j) \tag{3.4}$$

Το κριτήριο της μέγιστης σχετικότητας αναζητά ένα σύνολο S που μεγιστοποιεί το $V_{C,S}$ για το συγκεκριμένο πρόβλημα αναγνώρισης με κλάσεις που ανήκουν στο C.
Τελικά ο αλγόριθμος MRMR προσπαθεί να ικανοποιήσει και τα δύο παραπάνω κριτήρια μεγιστοποιώντας τον λόγο:

$$max_{S \subset \Omega} \frac{\sum_{i \in S} I(C,i)}{\frac{1}{|S|} \sum_{i,j \in S} I(i,j)} \tag{3.5}$$

Για την εύρεση της βέλτιστης λύσης, η πολυπλοκότητα του αλγορίθμου είναι $O(|\Omega|^{|S|})$ όπου $|\Omega|$ ο συνολικός αριθμός των χαρακτηριστικών στο πρόβλημα μας, καθώς χρειάζεται εξαντλητική αναζήτηση όλων των υποσυνόλων μεγέθους $|S|$. Ωστόσο στην πράξη κυρίως για λόγους πολυπλοκότητας υπολογίζουμε μια λύση που είναι κοντά στη βέλτιστη. Ο τρόπος είναι ο παρακάτω:

Αρχικά επιλέγουμε το 1ο χαρακτηριστικό του συνόλου, το οποίο είναι αυτό που μεγιστοποιεί το κριτήριο της μέγιστης σχετικότητας. Στη συνέχεια τα υπόλοιπα χαρακτηριστικά επιλέγονται με ακολουθητικό τρόπο, όμοια με την στρατηγική αναζήτησης FFS που περιγράψαμε προηγουμένως. Πιο συγκεκριμένα, αν S είναι το τρέχον σύνολο μας, τότε κάθε φορά προσθέτουμε το χαρακτηριστικό i για το οποίο μεγιστοποιείται ο λόγος:

$$max_{i \in \Omega - S} \frac{I(C,i)}{\frac{1}{|S|}\sum_{j \in S} I(i,j)} \tag{3.6}$$

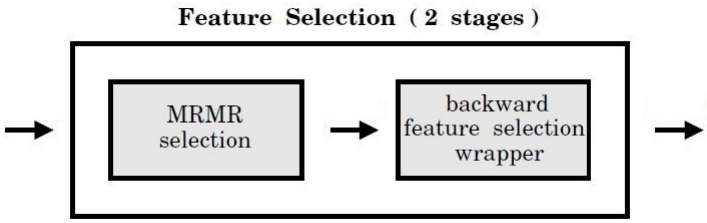
Η διαδικασία τερματίζει μόλις φτάσουμε στο επιθυμητό μέγεθος συνόλου. Χρειάζεται να πούμε ότι προκειμένου να χρησιμοποιήσουμε την συνάρτηση της αμοιβαίας πληροφορίας για διακριτές μεταβλητές όπως ορίστηκε παραπάνω, πρέπει αρχικά να διακριτοποιήσουμε όλα τα χαρακτηριστικά του προβλήματος.

3.3 Συνδυαστικός αλγόριθμος επιλογής χαρακτηριστικών

Όπως αναφέραμε και προηγουμένως, στους αλγόριθμους τύπου wrapper, η χρησιμότητα των χαρακτηριστικών αξιολογείται σύμφωνα με την ακρίβεια αναγνώρισης που πετυχαίνει το σύστημα μας χρησιμοποιώντας τα. Η διαδικασία αυτή είναι συνήθως χρονοβόρα, επειδή περιλαμβάνει εκπαίδευση του συστήματος για κάθε υποψήφιο υποσύνολο χαρακτηριστικών, ενώ επιπλέον αγνοείται η συσχέτιση και η αλληλοεπικάλυψη μεταξύ των χαρακτηριστικών που επιλέγονται. Από την άλλη, οι μέθοδοι τύπου filter λαμβάνουν υπόψιν τους κριτήρια αμοιβαίας πληροφορίας και στατιστικής συσχέτισης μεταξύ των χαρακτηριστικών, διατηρώντας παράλληλα χαμηλή πολυπλοκότητα που τους επιτρέπει να είναι απλές και γρήγορες στην υλοποίηση. Ωστόσο οι αλγόριθμοι φίλτρου αγνοούν τελείως τον τύπο του ταξινομητή που χρησιμοποιούμε κάτι που συχνά παίζει σημαντικό ρόλο στο τελικό αποτέλεσμα.

Προκύπτει λοιπόν με λογικό τρόπο η ιδέα ανάπτυξης ενός συνδυαστικού συστήματος επιλογής χαρακτηριστικών που θα χρησιμοποιεί και τις δύο παραπάνω κατηγορίες. Μια πιθανή λύση είναι αλγόριθμος δύο σταδίων, όπου στο πρώτο στάδιο προηγείται κάποια μέθοδος φίλτρου ενώ στο δεύτερο στάδιο εφαρμόζεται μέθοδος wrapper στο υποσύνολο που προέκυψε. Με αυτόν τον τρόπο συνδυάζονται τα θετικά σημεία και των δύο μεθόδων, ενώ παράλληλα μειώνεται σημαντικά η πολυπλοκότητα καθώς ο wrapper έχει πλέον ως είσοδο ένα αρκετά μικρότερο σύνολο από ασυσχέτιστα χαρακτηριστικά. Στα πειράματα μας, χρησιμοποιούμε στο πρώτο στάδιο τον αλγόριθμο MRMR ο οποίος επιλέγει περίπου το 1/3 των αρχικών χαρακτηριστικών, και στη συνέχεια έναν από τους αλγορίθμους FSS, BSS για να προκύψει το τελικό

διάνυσμα χαρακτηριστικών.



Σχήμα 3.2: Συνδυαστικός αλγόριθμος επιλογής 2 σταδίων [11]

Κεφάλαιο 4

Μελέτη ταξινομητών

Η επιλογή του κατάλληλου ταξινομητή αποτελεί το τελευταίο στάδιο στα προβλήματα της αναγνώρισης προτύπων, και είναι εξίσου σημαντική με την εξαγωγή χαρακτηριστικών. Το καταλληλότερο είδος ταξινομητή εξαρτάται γενικά από το πρόβλημα που έχουμε να αντιμετώπισουμε. Στην δική μας περίπτωση, πειραματιστήκαμε με διάφορους ταξινομητές (KNN, HMMs, GMMs, SVM), ωστόσο τα περισσότερα πειράματα έγιναν με βάση τους δύο τελευταίους. Στο κεφάλαιο αυτό θα περιγράψουμε τον τρόπο λειτουργίας των ταξινομητών GMM και SVM.

4.1 Ταξινομητής με χρήση μίγματος Γκαουσιανών (GMM)

Η ταξινόμηση με GMM (Gaussian mixture model) στηρίζεται στην ιδέα ότι η συνάρτηση πυκνότητας πιθανότητας μιας τυχαίας μεταβλητής μπορεί να προσεγγιστεί με ένα γραμμικό συνδυασμό από συναρτήσεις κανονικής κατανομής. Για κάθε κλάση συναισθήματος, υπολογίζεται ένα μοντέλο για να προσεγγίσει την συνάρτηση πυκνότητας πιθανότητας που ακολουθεί η μεταβλητή του διανύσματος χαρακτηριστικών. Στην συνέχεια κατηγοριοποιεί ένα νέο δείγμα στην κλάση της οποίας το μοντέλο του δίνει την μεγαλύτερη πιθανότητα. Παρακάτω περιγράφουμε πιο αναλυτικά τα βασικά σημεία του αλγόριθμου ταξινόμησης.

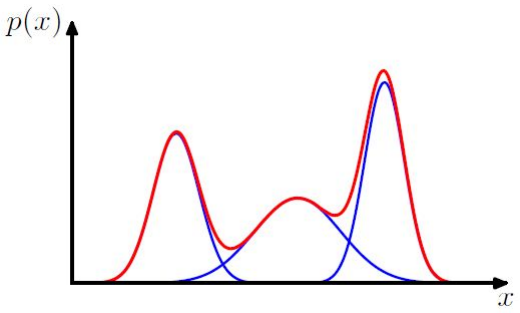
4.1.1 Μοντέλο μίγματος Γκαουσιανών

Ένα μοντέλο μίγματος M Γκαουσιανών δίνεται από την παρακάτω σχέση:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} w_i g(\mathbf{x}|\mu_i, \mathbf{\Sigma}_i), \tag{4.1}$$

όπου $\mathbf{x}$ είναι η D-διάστατη μεταβλητή, (π.χ το διάνυσμα χαρακτηριστικών), w_i, i=1,...,M είναι τα βάρη του γραμμικού συνδιασμού, και $g(\bullet|\mu_i, \Sigma_i)$ είναι οι επιμέρους Γκαουσιανές του μίγματος. Ο αναλυτικός τύπος μιας Γκαουσιανής D διαστάσεων είναι ο παρακάτω:

$$g(\mathbf{x}|\mu_i, \mathbf{\Sigma}_i) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_i|^{1/2}} e^{\{-\frac{1}{2}(\mathbf{x}-\mu_i)^T \mathbf{\Sigma}_i^{-1}(\mathbf{x}-\mu_i)\}}, \tag{4.2}$$



Σχήμα 4.1: Μίγμα Γκαουσιανών μιας μεταβλητής [49]

όπου μ_i είναι το διάνυσμα των μέσων τιμών και Σ_i είναι ο πίνακας συμμεταβλητότητας, ενώ τα βάρη του αθροίσματος πρέπει να ικανοποιούν την σχέση $\sum_{i=1}^{M} w_i = 1$. Για να ορισθεί πλήρως ένα μίγμα M Γκαουσιανών χρειάζεται να προσδιοριστούν οι παράμετροι $\lambda = \{w_i, \mu_i, \mathbf{\Sigma}_i\}, i = 1, ..., M$. Οι πίνακες συμμεταβλητότητας συχνά θεωρούνται διαγώνιοι υποθέτωντας ανεξαρτησία μεταξύ των D μεταβλητών του διανύσματος $\mathbf{x}$.

4.1.2 Υπολογισμός παραμέτρων του GMM

Με δεδομένα τα δείγματα εκπαίδευσης που αντιστοιχούν σε κάποια κλάση, επιθυμούμε να υπολογίσουμε τις παραμέτρους λ του μοντέλου GMM, έτσι ώστε η κατανομή που θα προκύψει να ταιριάζει όσο το δυνατόν περισσότερο με την κατανομή των δειγμάτων. Για τον υπολογισμό των παραμέτρων χρησιμοποιείται η διαδεδομένη μέθοδος της μέγιστης πιθανοφάνειας.

Μέθοδος μέγιστης πιθανοφάνειας

Έστω ότι έχουμε ένα σύνολο S από n δείγματα εκπαίδευσης $\mathbf{x}_1, ..., \mathbf{x}_n$ τα οποία ανήκουν σε κάποια κλάση. Θεωρούμε ότι τα δείγματα αυτά ακολουθούν μια κατανομή $p(\mathbf{x})$ με γνωστή παραμετρική μορφή, έτσι ώστε αυτή να προσδιορίζεται μοναδικά από ένα διάνυσμα παραμέτρων λ. Για να δείξουμε την εξάρτηση από το διάνυσμα παραμέτρων συμβολίζουμε την κατανομή με $p(\mathbf{x}\,|\lambda)$. Η μέθοδος της μέγιστης πιθανοφάνειας θεωρεί ότι το διάνυσμα παραμέτρων έχει σταθερή τιμή η οποία πρέπει να προσδιορισθεί. Ως καταλληλότερη τιμή για την παράμετρο λ θεωρείται η τιμή που μεγιστοποιεί την πιθανότητα να προέκυψαν τα δείγματα εκπαίδευσης που διαθέτουμε. Θεωρώντας ότι τα δείγματα προέκυψαν με τρόπο ανεξάρτητο μεταξύ τους, η συνολική πιθανότητα γράφεται ως εξής:

$$p(S\|\lambda) = \prod_{k=1}^{n} p(\mathbf{x}_k\|\lambda) \tag{4.3}$$

Επειδή η πιθανότητα αυτή εξαρτάται από το λ, ονομάζεται συνάρτηση πιθανοφάνειας

του λ για το σύνολο S των δειγμάτων. Ο λογάριθμος της πιθανοφάνειας γράφεται:

$$\ln p(S\|\lambda) = \sum_{k=1}^{n} \ln \left\{ \sum_{i=1}^{M} w_i g(\mathbf{x}_k|\mu_i, \mathbf{\Sigma}_i) \right\} \tag{4.4}$$

Αλγόριθμος EM (Expectation-Maximization)

Από τον προηγούμενο τύπο φαίνεται ότι για τον υπολογισμό των παραμέτρων που μεγιστοποιούν την πιθανοφάνεια δεν υπάρχει κάποια αναλυτική σχέση κλειστού τύπου. Μια επιλογή είναι η χρήση κλασσικών μεθόδων αριθμητικής ανάλυσης. Εναλλακτικά χρησιμοποιούμε τον γνωστό αλγόριθμο EM (expectation maximization). Η βασική ιδέα του αλγορίθμου είναι ότι ξεκινώντας με ένα αρχικό μοντέλο με παραμέτρους λ, προσπαθεί να υπολογίσει ένα νέο μοντέλο με παραμέτρους $\bar{\lambda}$, έτσι ώστε να ισχύει $p(S\,|\bar{\lambda}) \geq p(S\,|\lambda)$. Στη συνέχεια το νέο αυτό μοντέλο γίνεται το αρχικό μοντέλο για την επόμενη επανάληψη, και η διαδικασία συνεχίζεται μέχρις ότου ικανοποιηθεί κάποιο όριο σύγκλισης. Έτσι ο αλγόριθμος EM σε κάθε επανάληψη αποτελείται από δύο βασικά βήματα τα οποία εξασφαλίζουν αύξουσα πορεία στην συνάρτηση πιθανοφάνειας (βήματα E και M). Παρακάτω περιγράφουμε τα βασικά βήματα του αλγορίθμου.

1. Αρχικοποίηση των παραμέτρων μέσης τιμής μ_i, συμμεταβλητότητας $\mathbf{\Sigma}_i$ και βαρών w_i και υπολογισμός της αρχικής τιμής της πιθανοφάνειας. Αρχικοποίηση των παραμέτρων με χρήση του αλγορίθμου k-means δίνει συνήθως καλύτερα αποτελέσματα.

2. **Βήμα E**. Υπολογισμός των μεταβλητών "ανάθεσης ευθύνης"

$$\gamma_{ki} = \frac{w_i g(\mathbf{x}_k\|\mu_i, \mathbf{\Sigma}_i)}{\sum_{j=1}^{M} w_j g(\mathbf{x}_k\|\mu_j, \mathbf{\Sigma}_j)} \tag{4.5}$$

Διαισθητικά, η μεταβλητή γ_{ki} μετράει το ποσοστό "ευθύνης" της i-οστής γκαουσιανής για την δικαιολόγηση του δείγματος $\mathbf{x}_k$. Ισχύει ότι $\gamma_{ki} = p(\lambda_i\,|\mathbf{x}_k)$

3. **Βήμα M**. Υπολογισμός των νέων τιμών των παραμέτρων χρησιμοποιώντας τις τρέχουσες τιμές των μεταβλητών γ_{ki} που βρέθηκαν στο βήμα E.

$$\mu_i^{new} = \frac{1}{N_i} \sum_{k=1}^{n} \gamma_{ki} \mathbf{x}_k \tag{4.6}$$

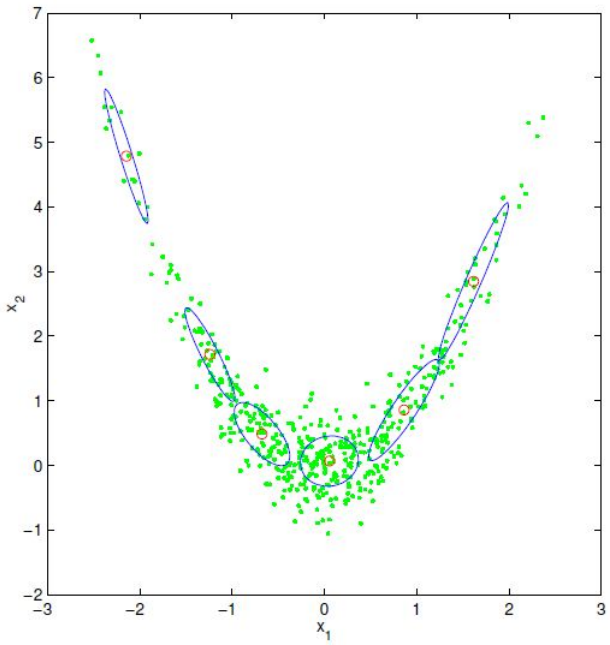
$$\Sigma_i^{new} = \frac{1}{N_i} \sum_{k=1}^{n} \gamma_{ki} (\mathbf{x}_k - \mu_i^{new})(\mathbf{x}_k - \mu_i^{new})^T \tag{4.7}$$

$$\pi_i^{new} = \frac{N_i}{n} \tag{4.8}$$

όπου $N_i = \sum_{k=1}^{n} \gamma_{ki}$.

4. Υπολογισμός του λογαρίθμου της πιθανοφάνειας με τις τρέχουσες τιμές των παραμέτρων $\lambda = \{\mu_i, \mathbf{\Sigma}_i, w_i\}$ και έλεγχος για το αν ικανοποιείται το κριτήριο σύγκλισης. Αν όχι, επιστροφή στο βήμα 2.

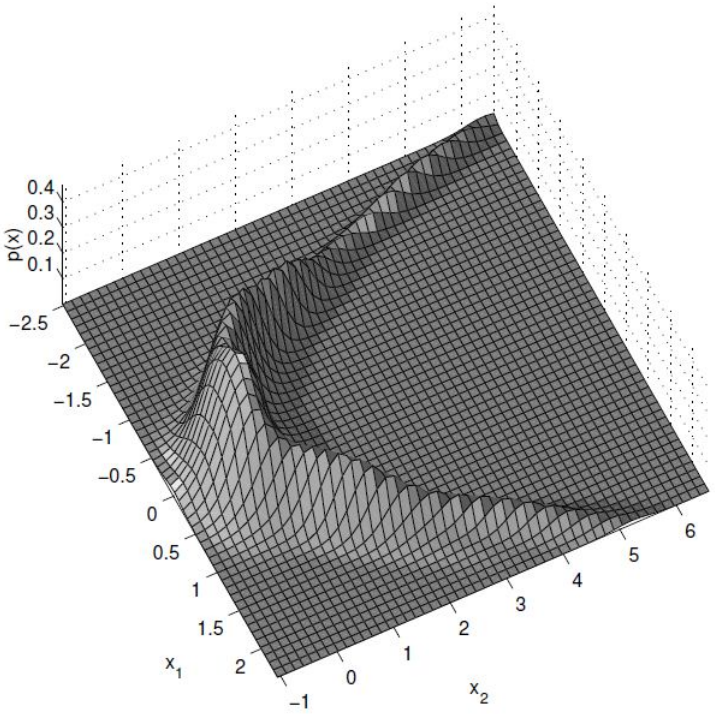
Στα σχήματα 4.2,4.3 φαίνεται το αποτέλεσμα της εφαρμογής του αλγορίθμου ΕΜ για την εύρεση του κατάλληλου μοντέλου μίγματος Γκαουσιανών για ένα σύνολο δειγμάτων εκπαίδευσης.



Σχήμα 4.2: Παράδειγμα υπολογισμού μίγματος Γκαουσιανών σε δεδομένα δύο διαστάσεων [50]

Σημειώνουμε ότι για την εύρεση των παραμέτρων λ του μοντέλου μας έχουμε υποθέσει ότι ο αριθμός M των Γκαουσιανών θεωρείται γνωστός. Το πλήθος των γκαουσιανών επηρεάζει σημαντικά την ακρίβεια με την οποία προσεγγίζεται η κατανομή των δειγμάτων και εξαρτάται κάθε φορά από το πρόβλημα. Απαιτείται λοιπόν πειραματισμός για την εύρεση του κατάλληλου πλήθους. Μεγάλο M οδηγεί σε υπερεκπαίδευση (over fitting) ενώ μικρό M δίνει ασθενή προσέγγιση της πραγματικής κατανομής. Ακόμα πολύ συχνά για χάρη μείωσης της πολυπλοκότητας των υπολογισμών, οι πίνακες συμμεταβλητότητας των Γκαουσιανών περιορίζονται στο να είναι διαγώνιοι. Ο διαγώνιος πίνακας συμμεταβλητότητας θεωρεί ανεξαρτησία μεταξύ των D μεταβλητών της Γκαουσιανής, κάτι που στην πράξη δεν ισχύει πάντοτε. Ωστόσο επειδή κατασκευάζουμε μίγμα από Γκαουσιανές μπορούμε με έναν αρκετό αριθμό από αυτές να προσεγγίσουμε έμμεσα τις αλληλεξαρτήσεις των μεταβλητών και να καλύψουμε σε ένα βαθμό την χαμένη πληροφορία που αγνοούμε.

Τέλος πρέπει να πούμε ότι όπως και άλλες μέθοδοι βελτιστοποίησης, ο αλγόριθμος ΕΜ δεν εγγυάται ότι βρίσκει το συνολικό μέγιστο της συνάρτησης πιθανοφάνειας. Στην γενική περίπτωση η λύση του συγκλίνει σε ένα από τα τοπικά μέγιστα της συνάρτησης το οποίο ελπίζουμε να είναι και από τα μεγαλύτερα. Η αρχικοποίηση των παραμέτρων του με χρήση του k-means βοηθάει προς αυτή την κατεύθυνση.



Σχήμα 4.3: Κατάλληλος αριθμός Γκαουσιανών μπορεί να προσεγγίσει οποιαδήποτε κατανομή [50]

4.2 Ταξινομητής SVM (Support Vector Machine)

Ο ταξινομητής SVM επινοήθηκε και διαμορφώθηκε στην τελική του μορφή από τον Vapnik [51] το 1995 και από τότε έχει χρησιμοποιηθεί ευρύτατα στις εφαρμογές της αναγνώρισης προτύπων. Ο SVM αναπαριστά τα διανύσματα χαρακτηριστικών των διάφορων δειγμάτων σαν σημεία στον m-διάστατο χώρο και στη συνέχεια προσπαθεί να βρεί μια επιφάνεια που θα τα διαχωρίζει βέλτιστα στις δύο διαφορετικές κλάσεις που ανήκουν. Βέλτιστη διαχωριστική επιφάνεια θεωρείται αυτή που μεγιστοποιεί την απόσταση της από το κοντινότερο σημείο οποιασδήποτε κλάσης, πετυχαίνοντας έτσι μέγιστο περιθώριο μεταξύ των δύο κλάσεων. Στη συνέχεια παρουσιάζουμε με περισσότερη λεπτομέρεια τον ταξινομητή.

4.2.1 Υπερεπίπεδο διαχωρισμού και μέγιστο περιθώριο

Όπως αναφέραμε, τα διανύσματα χαρακτηριστικών μήκους m που έχουμε εξάγει από τα δείγματα των διαφόρων κλάσεων αντιστοιχίζονται σε σημεία στον m-διάστατο χώρο: $\mathbf{x}_i \in \Re^m$. Η εξίσωση που περιγράφει ένα υπερεπίπεδο στον χώρο αυτό έχει την μορφή:

$$\mathbf{w}^T\mathbf{x} + b = 0, w \in \Re^m, b \in \Re, \tag{4.9}$$

όπου w είναι διάνυσμα κάθετο στο υπερεπίπεδο και b σταθερά.

Αρχικά υποθέτουμε ότι τα δείγματα ανήκουν σε δύο διαφορετικές κλάσεις και ότι τα σημεία στον m-διάστατο χώρο είναι γραμμικώς διαχωρίσιμα, υπάρχει δηλαδή ένα υπερεπίπεδο το οποίο να τα διαχωρίζει στις δύο αυτές κλάσεις. Στην πραγματικότητα υπάρχουν άπειρα τέτοια υπερεπίπεδα, ωστόσο εμείς επιλέγουμε αυτό που αφήνει το μέγιστο περιθώριο μεταξύ των δύο κλάσεων. Ο λόγος που θεωρούμε βέλτιστο αυτό το υπερεπίπεδο είναι ο εξής: Ο βασικός σκοπός του ταξινομητή είναι να μαθαίνει από παραδείγματα και στη συνέχεια να αναγνωρίζει την κλάση ενός νέου δείγματος εξετάζοντας την θέση του σε σχέση με αυτά της εκπαίδευσης. Αν επιλέγαμε ένα διαχωριστικό υπερεπίπεδο το οποίο απέχει μικρή απόσταση από τα σημεία κάποιας κλάσης, τότε θα υπήρχε μεγαλύτερη πιθανότητα ένα νέο δείγμα να βρεθεί κοντά στο σύνορο και να κατηγοριοποιηθεί λανθασμένα. Η ικανότητα ενός ταξινομητή να γενικεύει από τα δείγματα που εκπαιδεύτηκε και να κατηγοριοποιεί σωστά νέα άγνωστα δείγματα ονομάζεται ικανότητα γενίκευσης (generalisation ability) και εξαρτάται από το περιθώριο μεταξύ του διαχωριστικού επιπέδου και των δύο κλάσεων.

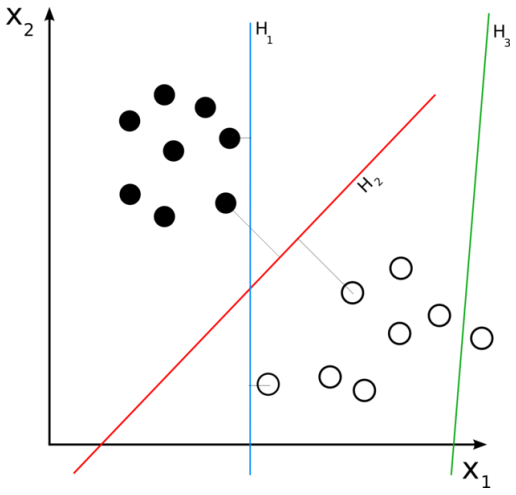
Προχωράμε με τους παρακάτω ορισμούς:

Διαχωρισιμότητα: Ένα σύνολο δειγμάτων εκπαίδευσης $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n) : \mathbf{x}_k \in \Re^m, y_k \in \{-1, +1\}\}$ (όπου το y_k υποδηλώνει την κλάση στην οποία ανήκει το σημείο $\mathbf{x}_k$) ονομάζεται γραμμικώς διαχωρίσιμο όταν υπάρχει ένα υπερεπίπεδο $\mathbf{w}^T\mathbf{x} + b = 0$ τέτοιο ώστε να ισχύουν:

$$\mathbf{w}^T\mathbf{x}_k + b \geq +1, \ \alpha\nu \ \ y_k = +1 \tag{4.10}$$

$$\mathbf{w}^T\mathbf{x}_k + b \leq -1, \ \alpha\nu \ \ y_k = -1 \tag{4.11}$$

Το υπερεπίπεδο που ορίζεται από τα $\mathbf{w}, b$ ονομάζεται διαχωριστικό.



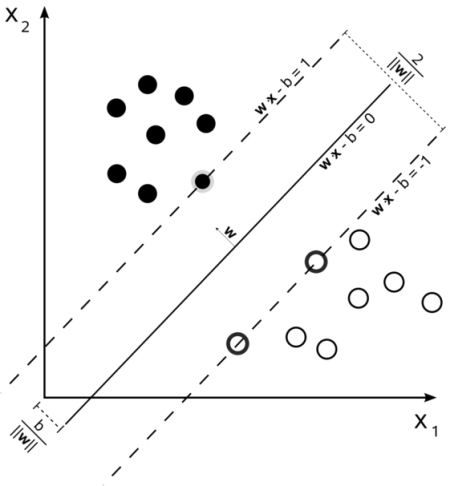
Σχήμα 4.4: Το υπερεπίπεδο H_3 δεν διαχωρίζει τις δύο κλάσεις. Το H_2 τις διαχωρίζει με μικρό όμως περιθώριο και τέλος το H_1 το πετυχαίνει με το μέγιστο περιθώριο [52].

Περιθώριο: Θεωρούμε το διαχωριστικό υπερεπίπεδο H που ορίζεται από την $\mathbf{w}^T\mathbf{x}+b = 0$. Η απόσταση ενός σημείου $\mathbf{x}_k$ από το επίπεδο αυτό ορίζεται ως:

$$\zeta_k(\mathbf{w}, b) = y_k(\mathbf{w}^T\mathbf{x} + b). \tag{4.12}$$

και το περιθώριο του συνόλου των σημείων S από το επίπεδο H, ορίζεται ως η απόσταση του H από το κοντινότερο σημείο:

$$\zeta_S(\mathbf{w}, b) = \min_{\mathbf{x}_k \in S} y_k(\mathbf{w}^T\mathbf{x} + b) \tag{4.13}$$



Σχήμα 4.5: Υπερεπίπεδο μέγιστου περιθωρίου για τον διαχωρισμό δειγμάτων δύο κλάσεων. Τα δείγματα που βρίσκονται πάνω στο περιθώριο ονομάζονται support vectors [52]

4.2.2 Εύρεση βέλτιστου υπερεπιπέδου

Για την κατασκευή λοιπόν του βέλτιστου διαχωριστικού υπερεπιπέδου απαιτούμε να ισχύουν ταυτόχρονα οι παρακάτω συνθήκες:

$$\max \ \ \zeta_S(\mathbf{w}, b) \tag{4.14}$$

$$\zeta_S(\mathbf{w}, b) \geq 1 \tag{4.15}$$

Και επειδή, όπως εύκολα δείχνεται γεωμετρικά, το εύρος του περιθωρίου ισούται με $1/\left\|\mathbf{w}\right\|$, για να μεγιστοποιήσουμε το περιθώριο αρκεί να ελαχιστοποιήσουμε το $\left\|\mathbf{w}\right\|$. Ισοδύναμα ελαχιστοποιούμε το $\left\|\mathbf{w}\right\|^2/2$ οπότε τελικά έχουμε να λύσουμε τις παρακάτω εξισώσεις που αποτελούν ένα πρόβλημα δευτεροβάθμιου προγραμματισμού (quadratic programming):

$$\min \ \ \frac{1}{2}\left\|\mathbf{w}\right\|^2 \tag{4.16}$$

$$y_k(\mathbf{w}^T\mathbf{x} + b) \geq 1, \quad k = 1, .., n \tag{4.17}$$

Εναλλακτικά μπορούμε να εκφράσουμε το πρόβλημα αν εισάγουμε την Langragian συνάρτηση:

$$L(\mathbf{w}, b, a) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{k=1}^{N} a_k\{y_k[\mathbf{w}^T\mathbf{x}_k + b] - 1\}, \quad a_k \geq 0, \quad k = 1, ..., n \tag{4.18}$$

και στη συνέχεια αναζητήσουμε το σημείο σέλλας (saddle point) της συνάρτησης:

$$\max_{\mathbf{a}} \min_{\mathbf{w},b} L(\mathbf{w}, b, \mathbf{a}) \tag{4.19}$$

Αρχικά για την ελαχιστοποίηση, απαιτώντας $\partial L/\partial \mathbf{w} = 0$ και $\partial L/\partial b = 0$, παίρνουμε τις παρακάτω σχέσεις:

$$\mathbf{w} = \sum_{k=1}^{n} a_k y_k \mathbf{x}_k, \quad \sum_{k=1}^{n} a_k y_k = 0 \tag{4.20}$$

και αντικαθιστώντας στην προηγούμενη σχέση, προκύπτει ότι πρέπει να μεγιστοποιήσουμε την συνάρτηση:

$$\max_{\mathbf{a}} L_D(\mathbf{a}) = -\frac{1}{2}\sum_{k,l=1}^{n} y_k y_l \mathbf{x}_k^T \mathbf{x}_l a_k a_l + \sum_{k=1}^{n} a_k, \quad \sum_{k=1}^{n} a_k y_k = 0 \tag{4.21}$$

4.2.3 Περίπτωση γραμμικώς μη διαχωρίσιμων δειγμάτων

Μέχρι τώρα έχουμε υποθέσει ότι οι δύο κλάσεις δειγμάτων είναι γραμμικά διαχωρίσιμες, κάτι που πολλές φορές όμως δεν συμβαίνει στην πράξη. Έτσι το 1995 οι Vapnik, Cortes [53], πρότειναν μια τροποποίηση του SVM η οποία αναζητά για ταξινομητή μέγιστου περιθωρίου επιτρέποντας όμως παράλληλα την λανθασμένη κατηγοριοποίηση κάποιων δειγμάτων. Η μέθοδος "χαλαρού" περιθωρίου (soft margin) όπως ονομάζεται, επιλέγει ένα υπερεπίπεδο το οποίο διαχωρίζει τα δείγματα όσο πιο ξεκάθαρα γίνεται προσπαθώντας παράλληλα να μεγιστοποιήσει την απόσταση του από τα κοντινότερα σωστά κατηγοριοποιημένα δείγματα. Πιο

συγκεκριμένα έχουμε:

Εισάγοντας μια θετική σταθερά ξ_k "χαλαρώνουμε" τους περιορισμούς και οι ανισοτικές σχέσεις μετατρέπονται στις παρακάτω:

$$y_k[\mathbf{w}^T\mathbf{x}_k + b] \geq 1 - \xi_k, \quad k = 1, ..., n. \tag{4.22}$$

Ψάχνουμε για την μικρότερη σταθερά χαλάρωσης η οποία ικανοποιεί την:

$$\xi_k = \max\{0, 1 - y_k[\mathbf{w}^T\mathbf{x}_k + b]\} \tag{4.23}$$

Οι τιμές των ξ_k προσδιορίζουν την θέση των $\mathbf{x}_k$ σε σχέση με το διαχωριστικό υπερεπίπεδο. Έτσι αν:

- $\xi_k \geq 1 : y_k[\mathbf{w}^T\mathbf{x}_k + b] < 0$, λανθασμένη κατηγοριοποίηση
- $0 < \xi_k < 1$: Το $\mathbf{x}_k$ κατηγοριοποιείται σωστά αλλά βρίσκεται εντός του περιθωρίου
- $\xi_k = 0$: Το $\mathbf{x}_k$ κατηγοριοποιείται σωστά και βρίσκεται έξω από το περιθώριο (ή στο σύνορο του).

Η αντικειμενική συνάρτηση που είχαμε προηγουμένως αυξάνεται τώρα κατά έναν όρο που την επιβαρύνει για τις μη μηδενικές τιμές των ξ_k και η βελτιστοποίηση μετατρέπεται σε μια ανταλλαγή μεταξύ μεγάλου περιθωρίου και μικρής επιβάρυνσης από τα λάθη. Η νέα αντικειμενική συνάρτηση είναι η παρακάτω:

$$\min_{\mathbf{w},\xi,b} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{k=1}^{n} \xi_k \right\} \tag{4.24}$$

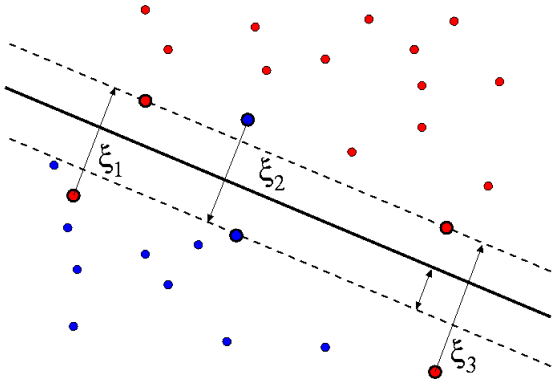
$$y_k[\mathbf{w}^T\mathbf{x}_k + b] \geq 1 - \xi_k, \quad \xi_k \geq 0 \quad k = 1, ..., n \tag{4.25}$$

Όπως φαίνεται από την παραπάνω σχέση όσο πιο λανθασμένα είναι κατηγοριοποιημένο ένα δείγμα τόσο περισσότερο επηρεάζει την τελική θέση του υπερεπιπέδου. Η επιρροή των λαθών στην θέση του υπερεπιπέδου γίνεται πιο χαλαρή με την μείωση της σταθεράς C.

4.2.4 Συναρτήσεις πυρήνα (Kernels)

Ο γραμμικός ταξινομητής SVM προσπαθεί να διαχωρίσει τα δείγματα με την χρήση ενός υπερεπιπέδου. Ο γραμμικός διαχωρισμός είναι αρκετά απλός και έχει καλές ιδιότητες γενίκευσης, ωστόσο κάποιες φορές μπορεί να μην ταιριάζει με την φύση του προβλήματος. Πολλές φορές τα δείγματα των δύο κλάσεων διαχωρίζονται καλύτερα από μη γραμμικές διαχωριστικές επιφάνειες. Η ιδέα για την λύση του προβλήματος είναι να αντιστοιχήσουμε με μια μη γραμμική συνάρτηση $\phi(\mathbf{x})$ τα σημεία $\mathbf{x}_k$ σε έναν χώρο περισσότερων διαστάσεων στον οποίο θα είναι γραμμικώς διαχωρίσιμα.

Επειδή στην επίλυση του προβλήματος εμφανίζονται εσωτερικά γινόμενα μεταξύ των αγνώστων, η μορφή της λύσης είναι ισοδύναμη με πριν. Έτσι εμείς δεν χρειάζεται να ξέρουμε την



Σχήμα 4.6: Παράδειγμα διαχωρισμού με soft margin [52]

μορφή της συνάρτησης $\phi(\mathbf{x})$, αλλά αρκεί να γνωρίζουμε το εσωτερικό γινόμενο $\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$, το οποίο το ονομάζουμε συνάρτηση πυρήνα (Kernel), $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$.
Οι ανισότητες κατηγοριοποίησης ξαναγράφονται ως εξής:

$$\mathbf{w}^T\phi(\mathbf{x}_k) + b \geq +1, \quad \alpha\nu \;\; y_k = +1 \tag{4.26}$$

$$\mathbf{w}^T\phi(\mathbf{x}_k) + b \leq -1, \quad \alpha\nu \;\; y_k = -1 \tag{4.27}$$

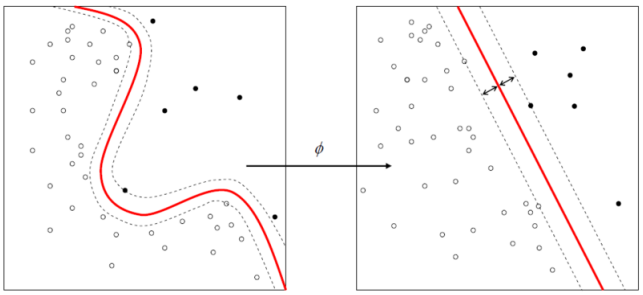
ενώ το νέο πρόβλημα βελιστοποίησης με χαλαρό περιθώριο ορίζεται από τις παρακάτω συνθήκες:

$$\min_{\mathbf{w},\xi,b}\left\{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{k=1}^{n}\xi_k\right\} \tag{4.28}$$

$$\mu\epsilon \quad y_k[\mathbf{w}^T\phi(\mathbf{x}_k) + b] \geq 1 - \xi_k, \quad \kappa\alpha\iota \quad \xi_k \geq 0 \quad k = 1, ..., n \tag{4.29}$$

Μερικές από τις πιο γνωστές συναρτήσεις Kernel είναι οι παρακάτω:

- Γραμμική: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$
- Πολυωνυμική βαθμού d: $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T\mathbf{x}_j + c)^d$



Σχήμα 4.7: Παράδειγμα διαχωρισμού με χρήση μη γραμμικής συνάρτησης Kernel [52]

- Γκαουσσιανή: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left[-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma}\right]$

Όπως παρατηρούμε, ο ταξινομητής SVM κατασκευάστηκε για να κάνει διαχωρισμό μεταξύ μόνο δύο κλάσεων δεδομένων. Σχεδιάστηκε δηλαδή για να είναι δυαδικός ταξινομητής. Αν θέλουμε επομένως να κάνουμε κατηγοριόποιηση δειγμάτων τα οποία ανήκουν σε $N > 2$ κλάσεις, θα πρέπει να σχεδιάσουμε ένα ιεραρχικό σύστημα το οποίο θα στηρίζεται σε συγκρίσεις ανά δύο μεταξύ των κλάσεων. Ένα παράδειγμα τέτοιου συστήματος είναι αυτό που για ένα νέο δείγμα, πραγματοποιεί όλες τις συγκρίσεις ανά δύο μεταξύ των κλάσεων και στη συνέχεια επιλέγει ως σωστή κλάση αυτή με τις περισσότερες νίκες (σχήμα majority-vote).

Κεφάλαιο 5

Πειράματα και αποτελέσματα

Στις προηγούμενες ενότητες αναπτύξαμε το θεωρητικό υπόβαθρο σχετικά με τα ακουστικά χαρακτηριστικά, τους αλγορίθμους επιλογής χαρακτηριστικών, και τους ταξινομητές που χρησιμοποιούμε στην έρευνα μας, με σκοπό την κατασκευή συστημάτων αναγνώρισης συναισθήματος μέσω φωνής. Σε αυτό το κεφάλαιο θα περιγράψουμε αναλυτικά τις διάφορες ερευνητικές προσπάθειες που έγιναν στα πλαίσια αυτής της διπλωματικής, μαζί με τα αποτελέσματα τους. Αρχικά παραθέτουμε κάποια στοιχεία για την βάση δεδομένων η οποία χρησιμοποιήθηκε, καθώς και για τον τρόπο με τον οποίο έγιναν τα πειράματα σε αυτή την βάση (πειραματικό πλαίσιο).

5.1 Βάσεις δεδομένων

Για την μελέτη της αναγνώρισης συναισθήματος μέσω φωνής έχουν αναπτυχθεί αρκετές βάσεις δεδομένων, οι οποίες όμως διαφέρουν ως προς την πειραματική διαδικασία συλλογής δεδομένων. Υπάρχουν τα παρακάτω βασικά θέματα που προκύπτουν κατά την κατασκευή μιας τέτοιας βάσης:

- Το πρώτο και βασικότερο αφορά στον τρόπο έκφρασης του συναισθήματος. Το συναίσθημα μπορεί να είναι αυθόρμητο, ή να προσομοιάζεται θεατρικά από ηθοποιούς.
- Ποιό θα πρέπει να είναι το επαρκές μέγεθος της βάσης σε σχέση με τον αριθμό των συναισθημάτων που απαιτούνται, και πόσοι ομιλητές πρέπει να συμμετέχουν.

Η εκφώνηση προκαθορισμένων προτάσεων από ηθοποιούς με προσποιητό συναίσθημα είναι η πιο συνήθης και εύκολα υλοποιήσιμη τεχνική που χρησιμοποιείται. Το πλεονέκτημα της μεθόδου αυτής είναι ότι οι συνθήκες δημιουργίας της βάσης είναι απόλυτα ελέγξιμες. Έτσι υπάρχει όσο το δυνατόν καλύτερη ποιότητα ηχογράφησης και επιλογή του επιθυμητού αριθμού συναισθημάτων που θα εκφωνηθούν.

Το βασικό μειονέκτημα των βάσεων προσποιητού λόγου είναι ότι αρκετοί ερευνητές αμφισβητούν την φυσικότητα και ομοιότητα των προσποιητών συναισθημάτων σε σχέση με τα γνήσια. Συχνά για παράδειγμα τα προσποιητά συναισθήματα μπορεί να είναι λίγο υπερβολικά και εντονότερα από τα αντίστοιχα αυθόρμητα. Ωστόσο η δημιουργία μιας βάσης δεδομένων με

στιγμιότυπα αυθόρμητου συναισθήματος είναι δύσκολη για αρκετούς λόγους (δυσκολίες ακριβούς χαρακτηρισμού συναισθήματος, δυσκολίες ηχογράφησης κλπ.), ενώ οι λιγοστές βάσεις που υπάρχουν δεν είναι γενικά διαθέσιμες στο κοινό για χρήση.

Στα πειράματα μας χρησιμοποιούμε την ευρέως διαδεδομένη βάση συναισθημάτων του Πανεπιστημίου του Βερολίνου (Emotional Berlin Database) [54], η οποία διατίθεται δωρεάν για χρήση. Η Berlin Database είναι βάση προσποιητού λόγου στην οποία εκφωνούνται στη γερμανική γλώσσα, προτάσεις με ουδέτερο περιεχόμενο από 10 ηθοποιούς (5 άνδρες και 5 γυναίκες) με διαφορετικά συναισθήματα κάθε φορά. Πιο συκγεκριμένα, εκφωνούνται 10 προτάσεις που χρησιμοποιούνται στην καθημερινότητα με 6 διαφορετικά συναισθήματα (θυμός, φόβος, χαρά, λύπη, απαρέσκεια, πλήξη) και το ουδέτερο. Ο λόγος που εκφωνούνται οι ίδιες προτάσεις για όλα τα συναισθήματα είναι για να γίνεται πιο εύκολη η σύκγριση μεταξύ συναισθημάτων και μεταξύ των ομιλητών, χωρίς να εμπλέκεται το περιεχόμενο της φράσης. Στη συνέχεια, οι εκφωνήσεις αυτές αξιολογήθηκαν ως προς το συναίσθημα που απεικονίζουν και ως προς την φυσικότητα τους και διατηρήθηκαν μόνο αυτές για τις οποίες η αναγνώριση συναισθήματος ήταν πάνω από 80% και η εκτιμώμενη φυσικότητα πάνω από 60%. Με τον τρόπο αυτό έγινε προσπάθεια τα προσποιητά συναισθήματα να καθρεφτίζουν όσο το δυνατόν καλύτερα τα αντίστοιχα γνήσια. Τελικά επιλέχτηκαν συνολικά περίπου 500 εκφωνήσεις.

Τέλος πρέπει να αναφέρουμε ότι η επιλογή της βάσης είναι πολύ σημαντική για την ποιότητα του συστήματος που θα φτιάξουμε. Και αυτό γιατί εκτός από το απαραίτητο υλικό που χρειαζόμασταν για να πειραματιστούμε, αποτελεί και το πρότυπο σύμφωνα με το οποίο εκπαδεύτηκε το σύστημα μας. Τα συναισθήματα που του διδάχθηκαν ως θυμός, χαρά, λύπη κλπ, προέρχονται από αυτή την βάση, και γι΄ αυτό θα πρέπει να είναι όσο το δυνατόν πιο ρεαλιστικά και αξιόπιστα.

5.2 Πειραματικό πλαίσιο

Στόχος μας είναι, δεδομένης της βάσης δεδομένων που διαθέτουμε, να εκπαιδεύσουμε ένα σύστημα το οποίο θα αναγνωρίζει με επιτυχία συναισθήματα από νέα άγνωστα δείγματα. Θέλουμε δηλαδή να φτιάξουμε ένα σύστημα αναγνώρισης με καλές ιδιότητες γενίκευσης. Για να γίνει αυτό χρησιμοποιώντας μόνο την βάση που διαθέτουμε, χρησιμοποιούμε συγκεκριμένο τρόπο αξιολόγησης του συστήματος. Συγκεκριμένα, διαμερίζουμε κάθε φορά την βάση σε 2 κύρια μέρη. Ένα μέρος χρησιμοποιείται μόνο για την εκπαίδευση (training data) και το υπόλοιπο χρησιμοποιείται για την αξιολόγηση (testing data). Τα αντίστοιχα ποσοστά είναι συνήθως 70-80% για εκπαίδευση και το υπόλοιπο για αξιολόγηση.

Στα δικά μας πειράματα, χρησιμοποιούμε την μέθοδο Leave-One-Out η οποία συνηθίζεται να χρησιμοποιείται σε συστήματα αναγνώρισης φωνής όπου συμμετέχουν διαφορετικοί εκφωνητές. Σύμφωνα με την μέθοδο αυτή μένουν τα δείγματα του ενός ομιλητή για τον ρόλο της αξιολόγησης, και οι υπόλοιποι χρησιμοποιούνται για την εκπαίδευση. Με τον τρόπο αυτό είναι φανερό πως μπορεί να κατασκευαστεί σύστημα με καλές ικανότητες γενίκευσης για αναγνώριση ανεξάρτητη από τον ομιλητή.

5.3 Ταξινόμηση με GMM

Στο πρώτο μας πείραμα επιχειρούμε να αναγνωρίσουμε τα έξι βασικά συναισθήματα (θυμός, φόβος, χαρά, λύπη, πλήξη, και ουδέτερο), χρησιμοποιώντας τον ταξινομητή GMM. Όπως αναφέραμε και προηγουμένως, για το κάθε συναίσθημα υπολογίζουμε ένα διαφορετικό μοντέλο μίγματος Γκαουσιανών, και στη συνέχεια κατηγοριοποιούμε ένα νέο δείγμα στην κλάση της οποίας το μοντέλο του δίνει την μεγαλύτερη πιθανοφάνεια. Στην συνέχεια περιγράφουμε λεπτομέρειες για τα διάφορα σημεία του συστήματος μας.

Λεπτομέρειες υλοποίησης του συστήματος

Για κάθε δείγμα φωνής που διαθέτουμε, κάνουμε εξαγωγή χαρακτηριστικών σε πλαίσια μικρής χρονικής διάρκειας (15-25 msec), αποκτώντας έτσι για κάθε ακουστικό χαρακτηριστικό μια κυματομορφή στο πεδίο του χρόνου. Συγκεκριμένα, τα χαρακτηριστικά MFCC υπολογίζονται σε πλαίσια διάρκειας 25 msec και με βήμα 15 msec (επικάλυψη 10 msec), ενώ όλα τα υπόλοιπα χαρακτηριστικά (Pitch, AM-FM, Glottal flow), υπολογίζονται σε μη επικαλυπτόμενα πλαίσια διάρκειας 15 msec. Για κάθε δείγμα φωνής λοιπόν, εξάγουμε ανά πλαίσιο τα παρακάτω χαρακτηριστικά:

- 12 MFCCs & 1η παράγωγος των MFCCs
- Θεμελιώδης συχνότητα (Pitch)
- 20 χαρακτηριστικά γλωττιδικού παλμού (Glottal-Flow) (18 στο πεδίο του χρόνου και 2 στο πεδίο της συχνότητας)
- 12 χαρακτηριστικά AM-FM

Έτσι, στην γενική περίπτωση που εξάγουμε συνολικά M διαφορετικά χαρακτηριστικά, αν ένα δείγμα φωνής αποτελείται από N πλαίσια, τότε θα προκύψουν N διανύσματα χαρακτηριστικών, μεγέθους M το καθένα. Το κάθε διάνυσμα χαρακτηριστικών αποτελεί μια παρατήρηση, και χρησιμοποιείται για την εκπαίδευση του αντίστοιχου μοντέλου μίγματος Γκαουσιανών.

Πειράματα

Στο πρώτο πείραμα, δοκιμάζουμε την ικανότητα των κλασσικών χαρακτηριστικών MFCC σε συνδυασμό με το pitch, για τον διαχωρισμό των 6 βασικών συναισθημάτων. Παράλληλα, για να επαληθεύσουμε την υπεροχή των συστημάτων που χρησιμοποιούν την γνώση του φύλου του ομιλητή, εκτελούμε το πείραμα με δύο διαφορετικούς τρόπους: Αρχικά φτιάχνουμε σύστημα, το οποίο χωρίς να λάβει υπόψιν του το φύλο, εκπαιδεύεται με τους 9 ομιλητές και αξιολογείται με τον 1 που απομένει, και στη συνέχεια, φτιάχνουμε σύστημα το οποίο εκπαιδεύεται και αξιολογείται ξεχωριστά για άνδρες και γυναίκες.

Από κάθε πλαίσιο λοιπόν, εξάγουμε 25 χαρακτηριστικά. Για το μοντέλο GMM χρησιμοποιούμε 2 έως 12 Γκαουσιανές με διαγώνιο πίνακα συμμεταβλητότητας. Καλύτερα αποτελέσματα παρατηρήθηκαν χρησιμοποιώντας 7 Γκαουσιανές για το κάθε μοντέλο. Τα αποτελέσματα συνοψίζονται στους πίνακες 5.1, 5.2 και 5.3.

	Χαρά	Θυμός	Φόβος	Λύπη	Πλήξη	Ουδέτερο
Χαρά	66.20%	15.49%	15.49%	0%	1.41%	1.41%
Θυμός	17.27%	77.27%	4.55%	0%	0%	0.91%
Φόβος	15.94%	4.35%	49.28%	11.59%	13.04%	5.80%
Λύπη	0%	0%	4.92%	90.16%	4.92%	0%
Πλήξη	1.25%	0%	12.50%	11.25%	55%	20%
Ουδέτερο	2.63%	0%	6.58%	3.95%	50%	36.84%

Πίνακας 5.1: Ποσοστά επιτυχίας σε όλους τους ομιλητές, με χρήση των $MFCCs$ και του *pitch*. Το συνολικό ποσοστό επιτυχίας είναι **62.74%**.

	Χαρά	Θυμός	Φόβος	Λύπη	Πλήξη	Ουδέτερο
Χαρά	61.36%	27.27%	4.55%	0%	2.27%	4.55%
Θυμός	30.19%	64.15%	5.66%	0%	0%	0%
Φόβος	0%	12.12%	66.67%	6.06%	9.09%	6.06%
Λύπη	0%	0%	11.11%	80.56%	8.33%	0%
Πλήξη	0%	0%	4.35%	0%	73.91%	21.74%
Ουδέτερο	0%	0%	0%	0%	40.54%	59.46%

Πίνακας 5.2: Ποσοστά επιτυχίας στις γυναίκες, με χρήση των $MFCCs$ και του *pitch*. Γυναικείο ποσοστό επιτυχίας: **67.47%**

	Χαρά	Θυμός	Φόβος	Λύπη	Πλήξη	Ουδέτερο
Χαρά	70.37%	25.93%	3.70%	0%	0%	0%
Θυμός	7.02%	82.45%	10.53%	0%	0%	0%
Φόβος	11.11%	8.33%	55.56%	2.78%	8.33%	13.89%
Λύπη	0%	0%	0%	100%	0%	0%
Πλήξη	0%	0%	5.88%	26.47%	58.82%	8.82%
Ουδέτερο	0%	0%	5.13%	15.38%	28.21%	51.28%

Πίνακας 5.3: Ποσοστά επιτυχίας στους άνδρες, με χρήση των $MFCCs$ και του *pitch*. Το ανδρικό ποσοστό επιτυχίας είναι ίσο με **69.27%**, και τέλος, το συνολικό ποσοστό επιτυχίας ίσο με **68.31%**.

Όπως επιβεβαιώνεται από το πρώτο πείραμα, το σύστημα που χρησιμοποιεί την γνώση του φύλου του ομιλητή είναι καλύτερο, και μάλιστα βελτιώνει το συνολικό ποσοστό αναγνώρισης περίπου κατά 6%.

Στο δεύτερο πείραμα δοκιμάζουμε την ικανότητα και των άλλων ομάδων ακουστικών χαρακτηριστικών στην αναγνώριση συναισθήματος. Η πιο συνηθισμένη μέθοδος για τον συνδυασμό διαφορετικών ομάδων χαρακτηριστικών είναι η ένωση τους σε ένα κοινό διάνυσμα χαρακτηριστικών και στην συνέχεια η επιλογή των καλύτερων από αυτά, με χρήση κάποιου αλγόριθμου επιλογής. Ωστόσο στην περίπτωση μας, λόγω του ότι, σε αντίθεση με τα MFCCs, υπολογίζουμε τα χαρακτηριστικά γλωττιδικού παλμού και τα χαρακτηριστικά AM-FM μόνο για τα πλαίσια στα οποία το pitch δεν είναι μηδενικό (έμφωνα πλαίσια), υπάρχει μια αναντιστοιχία η οποία μας απαγορεύει να δημιουργήσουμε κοινό διάνυσμα χαρακτηριστικών για κάθε πλαίσιο. Μια πιθανή λύση θα ήταν να υπολογίζαμε στατιστικά μεγέθη από την κυματομορφή του κάθε χαρακτηριστικού, όπως η μέση τιμή και η διασπορά, και να τα χρησιμοποιούσαμε αντί για την τιμή του χαρακτηριστικού σε κάθε πλαίσιο. Έτσι θα ήταν δυνατή η ένωση των διαφορετικών διανυσμάτων σε ένα τελικό, για κάθε δείγμα φωνής. Σε αυτή την περίπτωση όμως, λόγω του σχετικά μικρού αριθμού δειγμάτων που θα προκύψουν, δεν είναι εφικτή η αποτελεσματική εκπαίδευση του μίγματος των Γκαουσιανών.

Αποφασίζουμε έτσι να συνδυάσουμε την ισχύ των διαφορετικών χαρακτηριστικών σε επίπεδο απόφασης (decision level fusion). Αυτό σημαίνει ότι για την κατηγοριοποίηση ενός νέου δείγματος φωνής, συνυπολογίζονται τα αποτελέσματα που δίνουν οι διάφορες ομάδες συντελεστών. Συγκεκριμένα, αν σε ένα δείγμα φωνής, το μοντέλο που εκπαιδεύτηκε με την μια ομάδα χαρακτηριστικών για μια κλάση συναισθήματος δίνει πιθανοφάνεια P_1, και το μοντέλο που εκπαιδεύτηκε με την δεύτερη ομάδα δίνει πιθανοφάνεια P_2 για το ίδιο συναίσθημα, τότε η τελική πιθανότητα θα ισούται με:

$$P = aP_1 + (1 - a)P_2, \tag{5.1}$$

όπου a είναι ο συντελεστής που καθορίζει την βαρύτητα της κάθε ομάδας στην τελική απόφαση. Στο πείραμα μας, η πρώτη ομάδα αποτελείται από τους συντελεστές MFCC μαζί με το pitch, και η δεύτερη ομάδα από τα χαρακτηριστκά γλωττιδικού παλμού μαζί με τα AM-FM. Δοκιμάζοντας διάφορες τιμές για τον συντελεστή a, παρατηρήσαμε τα καλύτερα αποτελέσματα για την τιμή $\mathbf{0.75}$. Η 1η ομάδα των κλασσικών χαρακτηριστικών δηλαδή έχει βαρύτητα 75% και η δεύτερη ομάδα 25%. Στους πίνακες 5.4, 5.5 φαίνονται τα αποτελέσματα.

	Χαρά	Θυμός	Φόβος	Λύπη	Πλήξη	Ουδέτερο
Χαρά	52.27%	25%	13.64%	0%	2.27%	6.82%
Θυμός	22.64%	66.04%	7.55%	0%	1.89%	1.89%
Φόβος	0%	9.09%	75.76%	6.06%	6.06%	3.03%
Λύπη	0%	0%	8.33%	88.89%	2.78%	0%
Πλήξη	0%	0%	6.52%	0%	69.57%	23.91%
Ουδέτερο	0%	0%	0%	2.70%	18.92%	78.38%

Πίνακας 5.4: Ποσοστά επιτυχίας στις γυναίκες, με χρήση συνδυασμού των χαρακτηριστικών. Γυναικείο ποσοστό επιτυχίας: **70.68%**.

	Χαρά	Θυμός	Φόβος	Λύπη	Πλήξη	Ουδέτερο
Χαρά	74.07%	25.93%	0%	0%	0%	0%
Θυμός	7.02%	91.23%	1.75%	0%	0%	0%
Φόβος	13.89%	5.56%	58.33%	11.11%	0%	11.11%
Λύπη	0%	0%	0%	96%	0%	4%
Πλήξη	2.94%	2.94%	0%	35.30%	32.35%	26.47%
Ουδέτερο	5.13%	0%	2.56%	5.13%	28.21%	58.97%

Πίνακας 5.5: Ποσοστά επιτυχίας στους άνδρες, με χρήση συνδυασμού των χαρακτηριστικών. Το ανδρικό ποσοστό επιτυχίας είναι ίσο με **69.27%**, και τελικά, το συνολικό ποσοστό επιτυχίας ίσο με **70.02%**.

Παρατηρήσεις–Συμπεράσματα

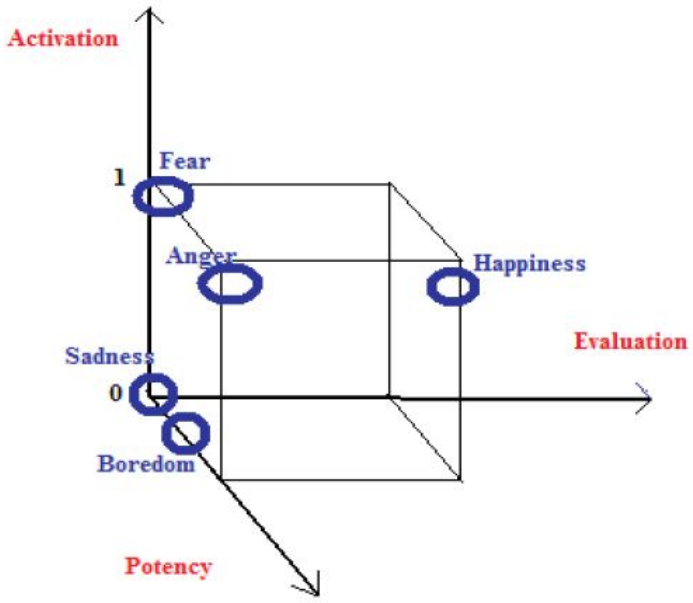
- Στο πρώτο πείραμα, χρησιμοποιώντας τα κλασσικά χαρακτηριστικά pitch και MFCC επαληθεύσαμε την ανωτερότητα των συστημάτων που λαμβάνουν υπόψιν τους το φύλο του ομιλητή. Επίσης είδαμε ότι χρησιμοποιώντας τον ταξινομητή GMM μπορούμε να πετύχουμε ποσοστά κοντά στο 70% για την συγκεκριμένη βάση συναισθημάτων.
- Στο δεύτερο πείραμα, είδαμε ότι τα χαρακτηριστικά γλωττιδικού παλμού και τα AM-FM χαρακτηριστικά συνέβαλαν σε μικρή βελτίωση (2%) του ποσοστού αναγνώρισης όταν συνδυάστηκαν με τις παραδοσιακές ομάδες χαρακτηριστικών. Αυτό δείχνει ότι συνδυασμένα με τα κλασσικά χαρακτηριστικά μπορούν να ενισχύσουν την ικανότητα του συστήματος. Μπορούμε επίσης να αναφέρουμε ότι από μόνα τους είχαν πιο φτωχή απόδοση (50%-55% επιτυχία) χρησιμοποιώντας τον ταξινομητή GMM με τις ίδιες παραμέτρους.

5.4 Ταξινόμηση με SVM με βάση τις διαστάσεις των συναισθημάτων

Εμπνευσμένοι από το ψυχολογικό μοντέλο ανάλυσης των συναισθημάτων σε διαστάσεις, επιχειρούμε να αναπτύξουμε ένα σύστημα που θα στηρίζεται στις διαφορετικές διαστάσεις των συναισθημάτων. Κύριος στόχος αυτής της προσπάθειας είναι η επαλήθευση της ισχύος αυτής της θεωρίας, και στην περίπτωση θετικής απάντησης, ο εντοπισμός των χαρακτηριστικών της φωνής που σχετίζονται με την κάθε διάσταση.

Αυτή την φορά χρισιμοποιούμε για το στάδιο της ταξινόμησης τον ταξινομητή SVM. Επειδή αυτός είναι σχεδιασμένος για να διαχωρίζει μόνο δύο διαφορετικές κατηγορίες, το τελικό σύστημα μας έχει ιεραρχική μορφή έτσι ώστε να μπορέσει να γίνει ταξινόμηση των 6 διαφορετικών συναισθημάτων. Όπως αναφέραμε, ο σχεδιασμός του συστήματος βασίζεται στις 3 βασικές διαστάσεις των συναισθημάτων valence, arousal, potency (ευχαρίστηση, διέγερση, έλεγχος).

Για την υλοποίηση του συστήματος, ορίζουμε δύο διακριτές στάθμες για την κάθε διάσταση, την *Υψηλή* και την *Χαμηλή* ({1,0}), και θεωρούμε πως κάθε συναίσθημα έχει μια από αυτές τις τιμές σε κάθε του διάσταση. Έτσι προκύπτουν οι παρακάτω ομαδοποιήσεις:



Σχήμα 5.1: Συναισθήματα σαν σημεία στον 3-διάστατο χώρο

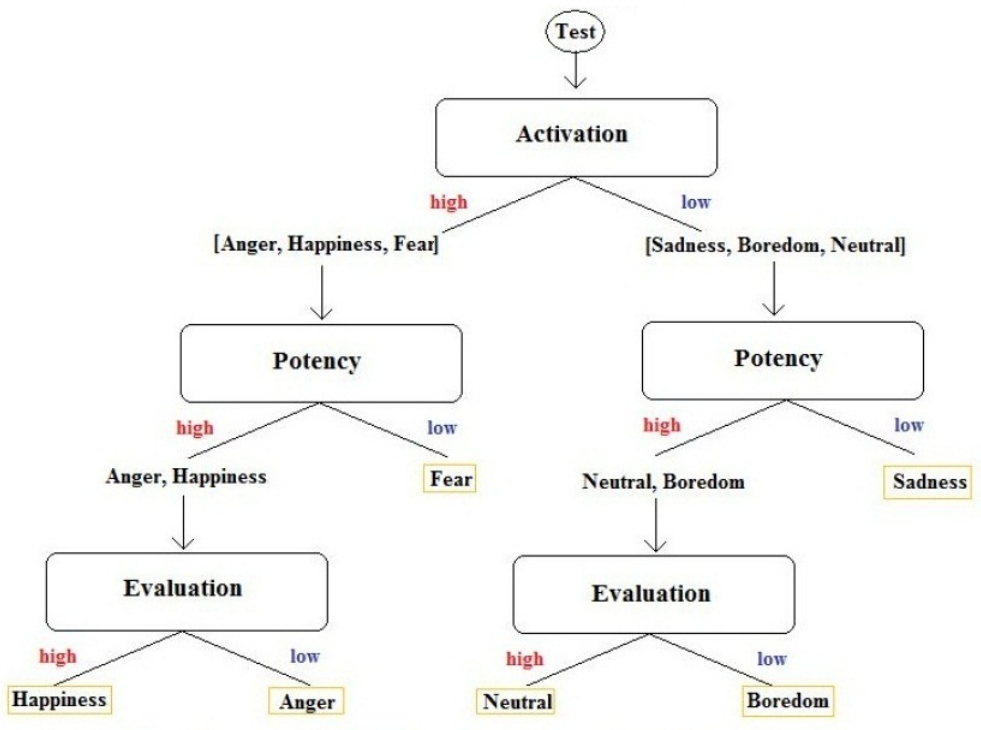
Διέγερση
- *Υψηλή Διέγερση*: Φόβος, Χαρά, Θυμός
- *Χαμηλή Διέγερση*: Πλήξη, Λύπη, Ουδέτερο

Ευχαρίστηση
- *Υψηλή Ευχαρίστηση*: Χαρά, (Ουδέτερο)
- *Χαμηλή Ευχαρίστηση*: Φόβος, Θυμός, Λύπη, Πλήξη

Έλεγχος
- *Υψηλός Έλεγχος*: Θυμός, Χαρά, (Ουδέτερο), (Πλήξη)
- *Χαμηλός Έλεγχος*: Φόβος, Λύπη

Μεσα σε παρένθεση γράφονται συναισθήματα τα οποία αν και στην πραγματικότητα δεν έχουν μια από τις τιμές 0 ή 1 σε κάποια διάσταση, για να μπορέσουμε να σχεδιάσουμε το σύστημα, τους δίνουμε την τιμή στην οποία βρίσκονται πιο κοντά. Έτσι π.χ. το συναίσθημα Πλήξη παίρνει την τιμή 1 στην διάσταση Έλεγχος.

Σύμφωνα με τις ομαδοποιήσεις αυτές (οι οποίες στηρίζονται σε ψυχολογικές αναλύσεις), κατασκευάστηκε το ιεραρχικό σύστημα ταξινόμησης που φαίνεται στο σχήμα 5.2.



Σχήμα 5.2: Ιεραρχικό σύστημα ταξινόμησης ανάλογα με τις τρείς βασικές διαστάσεις

Λεπτομέρειες υλοποίησης του συστήματος

Όπως φαίνεται από το σχήμα 5.2, χρειάζεται να εκπαιδεύσουμε 3 υποσυστήματα, ένα για κάθε βασική διάσταση, τα οποία θα κατηγοριοποιούν κάθε δείγμα του συνόλου αξιολόγησης σε μία από τις κλάσεις 1 (Υψηλό) και 0 (Χαμηλό). Έτσι ένα νέο άγνωστο δείγμα, θα ακολουθεί μία από τις πιθανές διαδρομές ανάλογα με τα αποτελέσματα των υποσυστημάτων, και θα καταλήγει σε ένα τελικό συναίσθημα.

Για κάθε δείγμα φωνής που διαθέτουμε, κάνουμε εξαγωγή χαρακτηριστικών σε πλαίσια βραχέος χρόνου της τάξης των 15-25 msec, και στην συνέχεια κάνουμε εκτιμήσεις για την μέση τιμή και την διασπορά του κάθε χαρακτηριστικού στο σύνολο της διάρκειας. Τα διαφορετικά είδη των χαρακτηριστικών που εξάγουμε ανά πλαίσιο ανάλυσης (frame) είναι τα παρακάτω:

Χαρακτηριστικά

- 12 MFCCs & 1η παράγωγος των MFCCs
- Θεμελιώδης συχνότητα (Pitch)
- Ενέργεια σήματος (Signal's energy)
- 20 χαρακτηριστικά γλωττιδικού παλμού (Glottal-Flow) (18 χρόνου και 2 συχνότητας)
- 12 χαρακτηριστικά AM-FM

Επιλογή χαρακτηριστικών

Αφού γίνει η εξαγωγή των 110 συνολικά χαρακτηριστικών, ακολουθεί το στάδιο της επιλογής των καταλληλότερων από αυτά για το κάθε υποσύστημα. Για την επιλογή των χαρακτηριστικών χρησιμοποιούμε το συνδυαστικό σχήμα των αλγορίθμων F-score και Forward Sequential Selection. Συγκεκριμένα, επιλέγονται αρχικά από τον αλγόριθμο φίλτρου F-score 40 χαρακτηριστικά και στην συνέχεια επιλέγονται από αυτά τα τελικά 12 χαρακτηριστικά με χρήση του FSS αλγορίθμου.

Τέλος αναφέρουμε ότι χρησιμοποιούμε την γνώση για το φύλο του ομιλητή, οπότε χρησιμοποιούμε δύο διαφορετικά συστήματα για άντρες και γυναίκες. Όπως φάνηκε και στο προηγούμενο πείραμα, η γνώση του φύλου βοηθάει στην βελτίωση της ταξινόμησης. Για αρχή, δείχνουμε παρακάτω τα τελικά αποτέλεσματα της ταξινόμησης στις βασικές κατηγορίες συναισθημάτων (συνυπολογίζοντας τα αποτελέσματα σε άντρες και γυναίκες).

Διαστάσεις	**Ποσοστό επιτυχίας**
Διέγερση	95.72%
Έλεγχος	84.84%
Ευχαρίστηση	70.69%

Πίνακας 5.6: Ποσοστό επιτυχίας για τις τρεις διαστάσεις συναισθημάτων

Όπως φαίνεται από τον προηγούμενο πίνακα, το ποσοστό επιτυχίας του υποσυστήματος για την διάσταση Διέγερση δεν είναι ικανοποιητικό. Αυτό μπορεί να οφείλεται είτε στην

ομαδοποίηση των συναισθημάτων (π.χ ουδέτερο στην ομάδα Υψηλό), είτε στον αρκετά μικρότερο αριθμό δειγμάτων της ομάδας 1 σε σχέση με την 0, είτε στην αδυναμία των εξαγόμενων χαρακτηριστικών για την συγκεκριμένο πρόβλημα κατηγοριοποίησης.
Τα τελικά αποτελέσματα που προκύπτουν χρησιμοποιώντας το σχήμα αυτό είναι τα παρακάτω:

Συναισθήματα	**Ποσοστό επιτυχίας ανδρ.**	**Ποσοστό επιτυχίας γυν.**
Χαρά	52%	63.64%
Θυμός	84.21%	67.92%
Φόβος-Άγχος	61.11%	54.55%
Λύπη	76%	69.44%
Πλήξη	67.65%	54.35%
Ουδέτερο	51.28%	75.67%

Πίνακας 5.7: Ποσοστά επιτυχίας συναισθημάτων για άνδρες και γυναίκες στο σύστημα που στηρίζεται στις ψυχολογικές διαστάσεις. Το ανδρικό ποσοστό επιτυχίας είναι ίσο με **67%**, το γυναικείο ποσοστό επιτυχίας ίσο με **64.26%**, και τέλος, το συνολικό ποσοστό επιτυχίας ισούται με **65.55%**.

Το τελικό αυτό ποσοστό αναγνώρισης δεν μπορεί να θεωρηθεί αρκετά ικανοποιητικό. Όπως παρατηρήσαμε και προηγουμένως, κύρια αδυναμία του είναι το υποσύστημα της διάστασης Ευχαρίστηση. Δομιμάζουμε λοιπόν να αντικαταστήσουμε στο τελευταίο στάδιο της ιεραρχίας το υποσύστημα της Ευχαρίστησης με δύο νέα υποστήματα. Το ένα θα είναι εξειδικευμένο για να διαχωρίζει χαρά και φόβο ενώ το άλλο πλήξη και ουδέτερο, οπότε το τελικό σύστημα γίνεται όπως φαίνεται στο σχήμα 5.3. Στους πίνακες 5.8-5.10 δείχνουμε αναλυτικά τα ποσοστά αναγνώρισης κάθε συναισθήματος στα διάφορα υποσυστήματα.

Συναισθήματα	Διέγερση	Έλεγχος	B-N / A-H
Χαρά	100%	100%	81.48%
Θυμός	100%	98.25%	80.7%
Φόβος-Άγχος	91.67%	66.67%	
Λύπη	100%	76%	
Πλήξη	97.06%	82.35%	79.41%
Ουδέτερο	89.74%	97.49%	77%

Πίνακας 5.8: Ποσοστά επιτυχίας συναισθημάτων για τους άνδρες στο βελτιωμένο σύστημα. Το υποσύστημα $B-N$ διαχωρίζει πλήξη από ουδέτερο, και το $A-H$ χαρά από θυμό.

Οπότε τελικά προκύπτει: Ανδρικό ποσοστό επιτυχίας = **75.22%**, Γυναικείο ποσοστό επιτυχίας = **69.08%**, Συνολικό ποσοστό επιτυχίας = **72.35%**.

Συναισθήματα	Διέγερση	Έλεγχος	B-N / A-H
Χαρά	95.45%	97.73%	82%
Θυμός	96.23%	94.34%	75.47%
Φόβος-Άγχος	93.94%	54.55%	
Λύπη	91.67%	77.78%	
Πλήξη	95.65%	78.26%	73.91%
Ουδέτερο	100%	91.89%	86.49%

Πίνακας 5.9: Ποσοστά επιτυχίας συναισθημάτων για τις γυναίκες στο βελτιωμένο σύστημα

Συναισθήματα	**Ποσοστό επιτυχίας ανδρ.**	**Ποσοστό επιτυχίας γυν.**
Χαρά	81.48%	75%
Θυμός	80.7%	71.7%
Φόβος-Άγχος	61.11%	54.55%
Λύπη	76%	69.44%
Πλήξη	76.47%	67.39%
Ουδέτερο	74.36%	72.97%

Πίνακας 5.10: Τελικά ποσοστά επιτυχίας συναισθημάτων για γυναίκες και άντρες στο βελτιωμένο σύστημα

Η βελτίωση που σημειώθηκε λοιπόν μετά την αντικατάσταση του συστήματος της Ευχαρίστησης είναι της τάξης 7%.

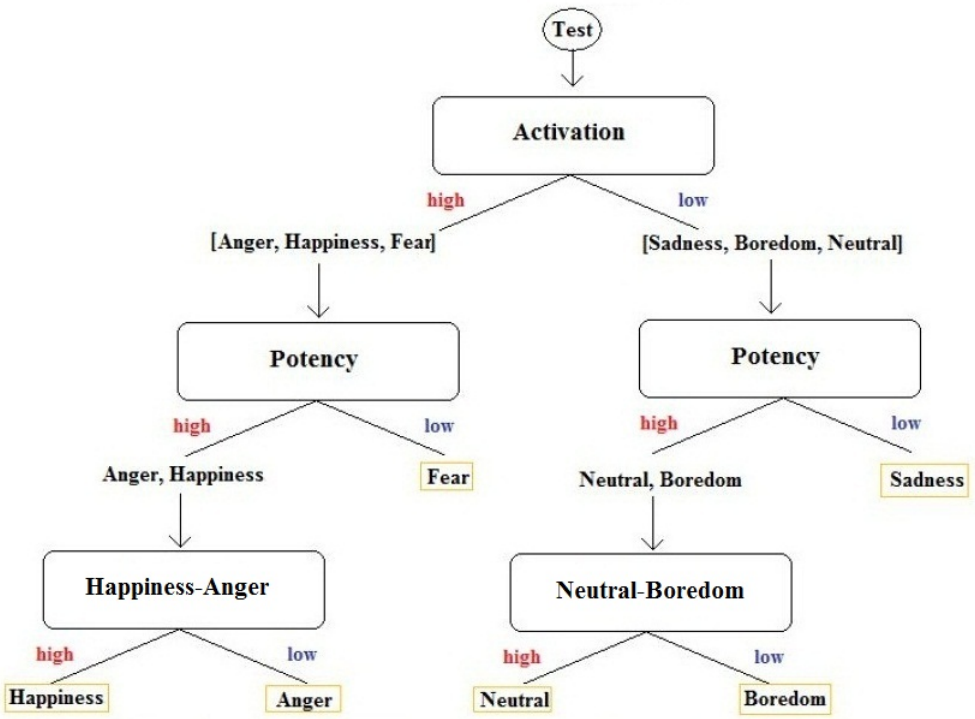
Παρατηρήσεις–Συμπεράσματα

- Επιχειρήσαμε να φτιάξουμε ένα σύστημα αναγνώρισης συναισθημάτων εμπνευσμένοι από το ψυχολογικό μοντέλο ανάλυσης τους στις βασικές διαστάσεις. Και το 1ο αλλά και το 2ο βελτιωμένο σύστημα στηρίζονται στην ομαδοποίηση των συναισθημάτων με βάση τις διαστάσεις Διέγερση και Έλεγχος. Λόγω της ιεραρχικής τους μορφής, η αναγνώριση ενός συναισθήματος είναι αρκετά αποδοτική από άποψη πολυπλοκότητας καθώς ένα νέο δείγμα περνάει από το πολύ 3 υποσυστήματα μέχρι να κατηγοριοποιηθεί. Επίσης αν κρίνουμε από το τελικό ποσοστό αναγνώρισης (72.35%), μπορούμε να πούμε πως επαληθεύεται σε κάποιο βαθμό η θεωρία των βασικών διαστάσεων, παρ'ολο που δεν είχαμε θεαματικά αποτελέσματα.

- Όσον αφορά στα χαρακτηριστικά που επιλέχθηκαν από τον αλγόριθμο επιλογής για κάθε διάσταση μπορούμε να πούμε τα εξής: Πράγματι, όπως φανταζόμασταν, παρατηρήθηκε προτίμηση σε συγκεκριμένες ομάδες χαρακτηριστικών για κάθε διάσταση. Έτσι για την διάσταση Διέγερση επιλέχθηκαν τα χαρακτηριστικά προσωδίας (Μέση τιμή και διασπο-

ρά του Πιτση και της ενέργειας σήματος), ενώ για την διάσταση Έλεγχος επιλέχθηκαν κυρίως Glottal Flow χαρακτηριστικά.

- Στην προσέγγιση αυτή, σε αντίθεση με την προηγούμενη, χρησιμοποιήσαμε κυρίως μέσες τιμές των χαρακτηριστικών, αντί για όλες τις τιμές τους ανά παράθυρο, και τα αποτελέσματα στην αναγνώριση ήταν παρόμοια. Συμπεραίνουμε λοιπόν ότι μπορούμε να μελετήσουμε το συναίσθημα μέσω φωνής σαν ένα φαινόμενο περισσότερο στατικό παρά δυναμικό, όπου ανάλογα με το συναίσθημα αλλάζουν κυρίως στατιστικά μεγέθη στα διάφορα ακουστικά χαρακτηριστικά.



Σχήμα 5.3: Βελτιωμένο ιεραρχικό σύστημα ταξινόμησης

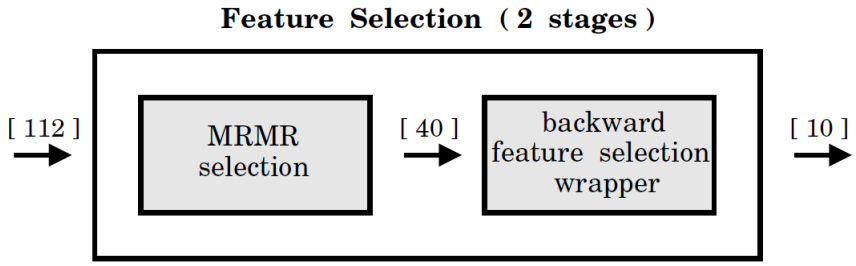
5.5 Ταξινόμηση με δυαδικό SVM σε ιεραρχικό σύστημα με χρήση σχήματος πλειοψηφίας (majority vote)

Η προηγούμενη προσπάθεια με το ιεραρχικό σύστημα ταξινόμησης στόχευε κυρίως στην επαλήθευση της ψυχολογικής θεωρίας περί διαστάσεων των συναισθημάτων, καθώς και στον εντοπισμό των υπεύθυνων-ρυθμιστικών χαρακτηριστικών της φωνής για την κάθε διάσταση. Όπως είδαμε στο 2ο βελτιωμένο σύστημα, η ενσωμάτωση δύο πιο εξειδικευμένων υποσυστημάτων βελτίωσε την συνολική απόδοση. Η παρατήρηση αυτή έρχεται να ενισχύσει την πεποίθηση μας ότι για το κάθε ζευγάρι συναισθημάτων υπάρχει διαφορετικό σύνολο χαρακτηριστικών το οποίο θα είναι το καταλληλότερο για τον διαχωρισμό τους.

Έτσι κατασκευάζουμε ένα νέο ιεραρχικό σύστημα το οποίο θα στηρίζεται σε συγκρίσεις ανά δύο των συναισθημάτων. Για κάθε ζεύγος θα εκπαιδεύεται και ένας διαφορετικός ταξινομητής SVM με χαρακτηριστικά που επιλέγονται από ένα νέο συνδυαστικό αλγόριθμο επιλογής. Στη συνέχεια περιγράφουμε πιο αναλυτικά τα διάφορα σημεία του νέου συστήματος.

Λεπτομέρειες υλοποίησης του συστήματος

Για την εκπαίδευση του κάθε υποσυστήματος, εξάγουμε χαρακτηριστικά από τα δείγματα εκπαίδευσης του αντίστοιχου ζεύγους συναισθημάτων. Τα ακουστικά χαρακτηριστικά που εξάγουμε στο κάθε σήμα είναι τα ίδια με το προηγούμενο πείραμα. Στην συνέχεια το διάνυσμα χαρακτηριστικών περνάει από έναν αλγόριθμο επιλογής 2 σταδίων, για να προκύψει τελικά ένα σύνολο από 10 χαρακτηριστικά, όπως φαίνεται στο σχήμα 5.4.

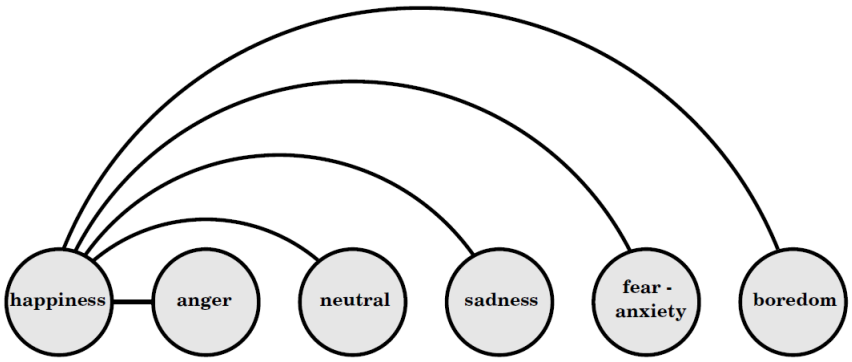


Σχήμα 5.4: Αλγόριθμος επιλογής χαρακτηριστικών 2 σταδίων για το τελικό πείραμα

Στο πρώτο στάδιο λοιπόν, εφαρμόζεται ο αλγόριθμος φίλτρου MRMR τον οποίο περιγράψαμε στην θεωρία, για να επιλεχθεί μια ομάδα από 40 χαρακτηριστικά. Τα χαρακτηριστικά αυτά είναι όσο το δυνατόν περισσότερο συσχετισμένα με το συγκεκριμένο ζεύγος συναισθημάτων, και επιπλέον έχουν μικρή αμοιβαία πληροφορία μεταξύ τους. Στην συνέχεια με χρήση του αλγορίθμου BSS (Backward Sequential Selection) καταλήγουμε σε 10 τελικά χαρακτηριστικά με τα οποία θα εκπαιδευτεί το υποσύστημα.

Επειδή έχουμε σαν πρόβλημα την αναγνώριση μεταξύ 6 βασικών συναισθημάτων, χρειαζόμαστε συνολικά 15 υποσυστήματα. Αφού εκπαιδευτούν και τα 15, για να ταξινομηθεί ένα καινούργιο δείγμα σε μια από τις 6 κατηγορίες γίνεται χρήση του σχήματος majority vote.

Το δείγμα δηλάδη περνάει σαν είσοδος και από τα 15 υποσυστήματα, σε καθένα από τα οποία επικρατεί ένα από τα δύο συναισθήματα. Στο τέλος επιλέγεται το συναίσθημα με τις περισσότερες επικρατήσεις. Σε περίπτωση ισοβαθμίας, επιλέγεται με τυχαίο τρόπο ένα από τα επικρατέστερα συναισθήματα.



Σχήμα 5.5: Υποσυστήματα για την αναγνώριση του συναισθήματος "Χαρά"

Πειράματα

Στην συνέχεια δείχνουμε τα αποτελέσματα διάφορων πειραμάτων στην μορφή του "πίνακα σύγχυσης" (Confusion Matrix). Όλα τα πειράματα έγιναν με την μέθοδο Leave-one-Speaker-out για να εξασφαλιστεί ανεξαρτησία από τον ομιλητή. Τα πειράματα στα οποία θεωρήθηκε γνωστό το φύλο του ομιλητή και έτσι εκπαιδεύτηκαν 2 συστήματα, ένα για κάθε φύλο, τα ονομάζουμε gender-dependent.

	Χαρά	Θυμός	Φόβος	Λύπη	Πλήξη	Ουδέτερο
Χαρά	74%	14%	10%	0%	1%	1%
Θυμός	10%	87%	2%	0%	1%	0%
Φόβος	7%	7%	78%	4%	3%	1%
Λύπη	0%	1%	7%	84%	6%	2%
Πλήξη	3%	0%	6%	7%	61%	23%
Ουδέτερο	0%	3%	3%	0%	12%	82%

Πίνακας 5.11: Ποσοστά επιτυχίας για gender dependent αναγνώριση με χρήση του Forward Selection αλγορίθμου για την επιλογή χαρακτηριστικών. Τελικό ποσοστό αναγνώρισης: **77.08%**.

Παρατηρώντας τα παραπάνω αποτέλεσματα μπορούμε να πούμε τα εξής: Καταρχήν

	Χαρά	Θυμός	Φόβος	Λύπη	Πλήξη	Ουδέτερο
Χαρά	73%	15%	9%	0%	1%	2%
Θυμός	12%	86%	1%	0%	1%	0%
Φόβος	5%	6%	82%	3%	3%	1%
Λύπη	0%	0%	5%	88%	5%	2%
Πλήξη	4%	0%	4%	6%	64%	22%
Ουδέτερο	1%	2%	3%	0%	10%	84%

Πίνακας 5.12: Ποσοστά επιτυχίας για gender dependent αναγνώριση με χρήση του Backward Selection αλγορίθμου. Τελικό ποσοστό αναγνώρισης: **79.71%**.

	Χαρά	Θυμός	Φόβος	Λύπη	Πλήξη	Ουδέτερο
Χαρά	81%	16%	2%	0%	1%	0%
Θυμός	7%	90%	2%	0%	1%	0%
Φόβος	5%	5%	81%	3%	6%	0%
Λύπη	0%	1%	3%	96%	0%	0%
Πλήξη	2%	0%	3%	7%	73%	15%
Ουδέτερο	0%	0%	2%	0%	8%	90%

Πίνακας 5.13: Ποσοστά επιτυχίας για gender dependent αναγνώριση με χρήση του συνδιαστικού αλγορίθμου επιλογής MRMR+Backward Selection. Τελικό ποσοστό αναγνώρισης: **85.18%**.

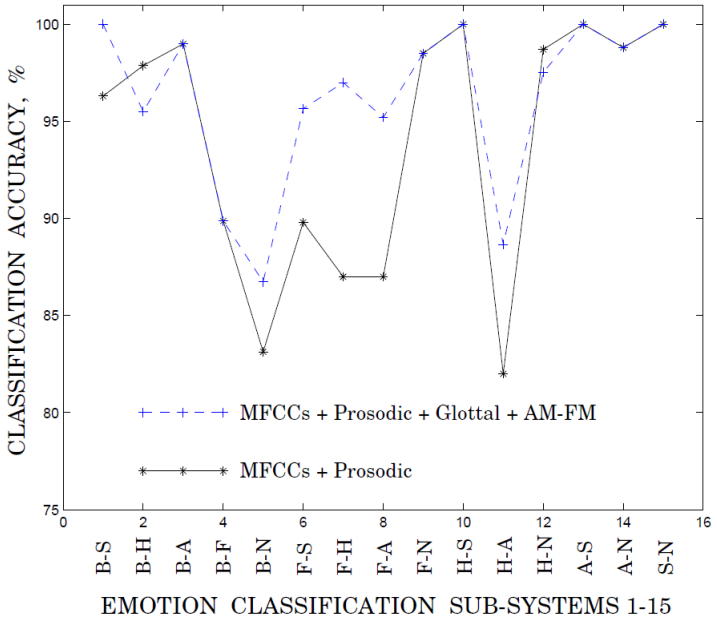
	Χαρά	Θυμός	Φόβος	Λύπη	Πλήξη	Ουδέτερο
Χαρά	77%	17%	3%	0%	3%	0%
Θυμός	11%	85%	4%	0%	0%	0%
Φόβος	7%	5%	74%	5%	7%	2%
Λύπη	0%	0%	3%	86%	6%	5%
Πλήξη	2%	0%	10%	6%	71%	11%
Ουδέτερο	2%	1%	3%	1%	9%	84%

Πίνακας 5.14: Ποσοστά επιτυχίας για gender independent αναγνώριση με χρήση του συνδιαστικού αλγορίθμου επιλογής MRMR+Backward Selection. Τελικό ποσοστό αναγνώρισης: **80.09**.

φαίνεται, για άλλη μια φορά, η ανωτερότητα της gender dependent προσέγγισης έναντι της gender independent. Ξεκάθαρα, η εκ των προτέρων γνώση του φύλου του ομιλητή βελτιώνει σημαντικά την απόδοση του αλγορίθμου. Επίσης παρατηρούμε σταδιακή βελτίωση των αποτελεσμάτων, αρχικά εφαρμόζοντας BSS αντί για FSS και στην συνέχεια χρησιμοποιώντας τον αλγόριθμο επιλογής 2 σταδίων MRMR+BSS.

Όσον αφορά στα ποσοστά αναγνώρισης, η καλύτερη εκδοχή του συστήματος μας φτάνει μέχρι και 85.18%. Το ποσοστό αυτό είναι συγκρίσιμο τόσο με τα state-of-the-art αποτελέσματα σε αυτή την βάση, όσο και με το ποσοστό αναγνώρισης που έχει σημειωθεί από ανθρώπους (84.3%) σε σχετική έρευνα [55].

Στο παρακάτω σχήμα δείχνουμε αποτελέσματα για την απόδοση των διάφορων υποσυστημάτων, ξεκινώντας από διαφορετικά αρχικά διανύσματα χαρακτηριστικών. Όπως φαίνεται, η προσθήκη των Glottal Flow και των AM-FM χαρακτηριστικών οδηγεί σε βελτιωμένα αποτελέσματα. Σε μερικά υποσυστήματα η χρήση μόνο των MFCCs και των χαρακτηριστικών προσωδίας είναι προτιμότερη. Αυτό δείχνει ότι τα χαρακτηριστικά αυτά είναι αρκετά αντιπροσωπευτικά από μόνα τους για να διαχωρίσουμε τα συγκεκριμένα ζεύγη συναισθημάτων.



Σχήμα 5.6: Ποσοστά αναγνώρισης των 15 υποσυστημάτων στα δείγματα ομιλητών γένους θηλυκού για δύο διαφορετικά αρχικά σύνολα χαρακτηριστικών. Στο ένα χρησιμοποιούνται τα κλασσικά χαρακτηριστικά MFCCs + prosodic ενώ στο άλλο προστίθενται και Glottal flow + AM-FM χαρακτηριστικά. Τα συναισθήματα συμβολίζονται με το αρχικό τους γράμμα στα αγγλικά.

Τέλος, όσον αφορά στα είδη των χαρακτηριστικών που επιλέχθηκαν από τον προτεινόμενο αλγόριθμο, μπορούμε να πούμε ότι πράγματι επιλέχθηκαν διαφορετικοί συνδιασμοί χαρακτηριστικών για το κάθε υποσύστημα, κάτι που ενισχύει την προσέγγιση μας. Για παράδειγμα αναφέρουμε πως όταν μια από τις δύο κλάσεις ήταν ο φόβος, επιλέχθηκαν περισσότερο Glottal flow χαρακτηριστικά.

Παρατηρήσεις-Συμπεράσματα

- Στο τελευταίο αυτό πείραμα, προτείναμε ένα ιεραρχικό σύστημα αναγνώρισης το οποίο στηρίζεται στην διαχωριστική ικανότητα των ειδικά εκπαιδευμένων υποσυστημάτων του. Τα υποσυστήματα αυτά χρησιμοποιούν τα καταλληλότερα από τα διάφορα είδη χαρακτηριστικά μέσω ενός νέου σχήματος επιλογής 2 σταδίων. Τέλος εκμεταλλευτήκαμε την γνώση του φύλου του ομιλητή για να πετύχουμε καλύτερα αποτελέσματα.

- Το τελικό μας σύστημα (gender dependent / MRMR+BSS) πέτυχε συνολικό ποσοστό επιτυχίας 85.18% το οποίο είναι συγκρίσιμο με τα state-of-the-art αποτελέσματα στην έρευνα της αναγνώρισης συναισθήματος μέσω φωνής, χρησιμοποιώντας μάλιστα σχετικά μικρό αριθμό χαρακτηριστικών.

- Και τα Glottal Flow αλλά και τα AM-FM χαρακτηριστικά επιλέχθηκαν από το σχήμα 2 σταδίων που χρησιμοποιήθηκε και βελτίωσαν τα ποσοστά επιτυχίας σε σχέση με την παραδοσιακή χρήση των MFCCs + prosodic χαρακτηριστικών.

- Το τελευταίο αυτό σύστημα είναι απλό στον σχεδιασμό, και κάθε νέο υποψήφιο για αναγνώριση δείγμα περνάει σαν είσοδος στα 15 διαφορετικά υποσυστήματα. Σε σχέση με το προηγούμενο πείραμα είναι λίγο περισσότερο περίπλοκο, αλλά δίνει αισθητά καλύτερα αποτελέσματα. Καταφέραμε έτσι, χρησιμοποιώντας απλούς δυαδικούς γραμμικούς ταξινομητές SVM να φτιάξουμε ένα αποδοτικό ιεραρχικό σύστημα.

Κεφάλαιο 6

Συμπεράσματα και μελλοντικές βελτιώσεις

Ολοκληρώνοντας την διπλωματική αυτή εργασία, συνοψίζουμε τα συμπεράσματα που προέκυψαν από τα αποτελέσματα των διαφόρων πειραμάτων, και αναφέρουμε πιθανές μελλοντικές αλλαγές που θα μπορούσαν να βελτιώσουν την έρευνα μας.

6.1 Συμπεράσματα

Στην διπλωματική μας εργασία ασχοληθήκαμε με την εξαγωγή διάφορων ομάδων ακουστικών χαρακτηριστικών, καθώς και με την δοκιμή διαφόρων αλγόριθμων επιλογής χαρακτηριστικών και ταξινομητών, με σκοπό την αναγνώριση συναισθήματος μέσω φωνής. Η έρευνα μας συνοψίζεται στα τρία βασικά πειράματα που περιγράψαμε στην προηγούμενη ενότητα.

Στο πρώτο πείραμα, με χρήση του ταξινομητή GMM επαληθεύτηκε ότι η εκ των προτέρων γνώση του φύλου του ομιλητή βοηθάει αισθητά στην αναγνώριση συναισθήματος. Επίσης διαπιστώθηκε ότι οι κατηγορίες χαρακτηριστικών AM-FM και Glottal flow, λειτουργούν ενισχυτικά στην αναγνώριση όταν συνδυαστούν με τις κλασσικές κατηγορίες. Στο δεύτερο πείραμα έγινε χρήση του ταξινομητή SVM σε ιεραρχικό σύστημα βασισμένο στις ψυχολογικές διαστάσεις των συναισθημάτων. Επαληθεύτηκε ως ένα βαθμό η θεωρία των ψυχολογικών διαστάσεων και αναζητήθηκαν τα σχετικά με την κάθε διάσταση χαρακτηριστικά. Τέλος, στο τρίτο πείραμα κατασκευάστηκε σύστημα που στηρίζεται στην ιδέα ότι το καταλληλότερο σύνολο χαρακτριστικών πρέπει να αναζητηθεί για κάθε ζεύγος συναισθημάτων ξεχωριστά. Για τον σκοπό αυτό δοκιμάστηκαν διάφορα σχήματα αλγορίθμων επιλογής και βρέθηκε ότι με τον συνδυαστικό αλγόριθμο επιλογής παρατηρήθηκαν τα καλύτερα αποτελέσματα. Στο τελευταίο αυτό πείραμα τα ποσοστά επιτυχίας είναι συγκρίσιμα με τα καλύτερα ερευνητικά αποτελέσματα στην αναγνώριση συναισθήματος ανεξάρτητης από τον ομιλητή, χρησιμοποιώντας μάλιστα αρκετά μικρότερο αριθμό χαρακτηριστικών.

6.2 Πιθανές μελλοντικές βελτιώσεις στην έρευνα μας

Όπως σε κάθε ερευνητική προσπάθεια, έτσι και στην δική μας δεν υπάρχει τέλος. Υπάρχουν πάντα ιδέες προς εφαρμογή στο άμεσο μέλλον και πιθανές μελλοντικές επεκτάσεις για τις οποίες υπάρχει ελπίδα ότι θα φέρουν βελτίωση στα αποτελέσματα. Οι κυριότερες μελλοντικές επεκτάσεις που σκοπεύουμε να γίνουν είναι οι παρακάτω:

- Στην καλύτερη προσπάθεια της παρούσας έρευνας, η πληροφορία που εξάγουμε από τα ακουστικά χαρακτηριστικά περιορίζεται στην μέση τιμή και την διασπορά της κυματομορφής τους. Θα έχει ιδιαίτερο και ουσιαστικό ενδιαφέρον να προσπαθήσουμε να χρησιμοποιήσουμε περισσότερη πληροφορία για την μορφή της καμπύλης του κάθε χαρακτηριστικού στο χρόνο. Πιστεύουμε ότι ενδεχομένως έως τώρα να μην εκμεταλλευόμαστε στο μέγιστο βαθμό την πληροφορία που μας δίνει η μεγάλη ποικιλία χαρακτηριστικών που εξάγουμε από τα δείγματα φωνής.

- Χρειάζεται να γίνει προσπάθεια για υλοποίηση συστήματος που θα αναγνωρίζει αυτόματα και επιτυχώς το φύλο του ομιλητή, καθώς η έρευνα μας στηρίζεται κυρίως σε gender-dependent συστήματα. Αν και υπάρχουν διαθέσιμα τέτοια συστήματα από άλλες ερευνητικές προσπάθειες, με την ποικιλία των χαρακτηριστικών που εξάγουμε από την φωνή, μπορούμε να δοκιμάσουμε να πετύχουμε ακόμα μεγαλύτερη απόδοση.

- Είναι απαραίτητος ο πειραματισμός σε παραπάνω από μια βάσεις δεδομένων για να αποκτήσει μεγαλύτερη αξιοπιστία η έρευνα. Τόσο ο αριθμός των διαφορετικών βάσεων που έχει δοκιμαστεί ένα σύστημα, όσο και το μέγεθος τους είναι πολύ σημαντικά κριτήρια που ενισχύουν την δύναμη του.
  Τέλος, δεν πρέπει να ξεχνάμε ότι η βάση στην οποία πειραματιζόμαστε έως τώρα (Berlin Database) περιέχει προσποιητό λόγο εκφωνημένο από επαγγελματίες ηθοποιούς. Θα είχε ιδιαίτερο ενδιαφέρον να δοκιμάσουμε τα πειράματα μας σε βάσεις με αυθόρμητο λόγο που θα περιέχει και περισσότερο ρεαλιστικά στιγμιότυπα των συναισθημάτων.

Bibliography

[1] X. Huahu, Y. Jian, and G. Jue, "Application of speech emotion recognition in intelligent household robot," *Artificial Intelligence and Computational Intelligence (AICI) 2010 International Conference*, vol. 1, pp. 537–541, 2010.

[2] A. Chitu, M. Vulpen, P. Takapoui, and L. Rothkrantz, "Building a dutch multimodal corpus for emotion recognition," in *LREC 2008, Workshop on Corpora for Research on Emotion and Affect*, 2008, pp. 53–56.

[3] R. Fernandez and R. Picard, "Signal processing for recognition of human frustration," *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 6, pp. 3773–3776, 1998.

[4] A. Razak, M. Yusof, and R. Komiya, "Emotion recognition in speech using a fuzzy approach," in *International Conference on Intelligent Knowledge Systems(IKS), Assos, Troy, Turkey*, 2004.

[5] D. J. France, R. Shiavi, S. Silverman, M. Silverman, and D. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on biomedical engineering*, vol. 7, pp. 829–837, 2000.

[6] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech," *IEEE transactions on biomedical engineering*, vol. 55, pp. 96–107, 2008.

[7] A. Mpoutri, "Organization and structure of emotions (in greek)," in *http* : $//www.positiveemotions.gr/index.php?option = com_content task = viewid = 28 Itemid = 55$, 2005.

[8] R. Plutchik, "A general psychoevolutionary theory of emotion," *Theories of emotion, New York: Academic*, vol. 1, pp. 3–33, 1980.

[9] E. Jee, Y. Cheong, C. Kim, D. Kwon, and H. Kobayashi, "Sound production for the emotional expression of socially interactive robots," *Advances in Human-Robot Interaction*, p. 342, 2009.

[10] B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," vol. 90, pp. 1415–1423, 2010.

[11] P. Giannoulis and G. Potamianos, "A hierarchical approach with feature selection for emotion recognition from speech," *LREC Instanbul*, 2012.

[12] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," *Proc. Int. Conf. Speech Prosody, Dressen, Germany*, 2006.

[13] P. Maragos, *Speech Processing Recognition (greek).* NTUA, Athens, 2002.

[14] G. Fant, "Acoustic theory of speech production," in *Mouton, The Hague*, 1960.

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

[16] M. Bulut, S. Lee, and S. Narayanan, "Recognition for synthesis: Automatic parameter selection for resynthesis of emotional speech from neutral speech," 2008.

[17] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," *Proc. ICSLP, Philadelphia, PA, USA*, pp. 1970–1973, 1996.

[18] V. Petrushin, "Emotion recognition agents in real world," in *AAAI Fall Symposium on Socially Intelligent Agents: Human in the Loop*, 2000.

[19] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features and methods," *Elsevier Speech Communication*, vol. 48, pp. 1162–1181, 2006.

[20] D. Talkin, "A robust algorithm for pitch tracking (rapt)," in *Speech Coding and Synthesis.* Elsevier, 1995, ch. 14, pp. 495–518.

[21] A. Paeschke, "Global trend of fundamental frequency in emotional speech," *ISCA - Speech Prosody, Nara, Japan (March 2004)*, vol. 18, pp. 671–674, 2004.

[22] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, pp. 32–80, 2001.

[23] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information," *Proc. European Signal Processing Conf. (EUSIPCO04)*, vol. 1, pp. 341–344, 2004.

[24] ——, "Emotional speech classification using gaussian mixture models and the sequential floating forward selection algorithm," *Proc. Int. Conf. Multimedia and Expo (ICME04)*, 2005.

[25] D. Cairns and J. Hansen, "Nonlinear analysis and classification of speech under stressed conditions," *J. Acoust. Soc. Am.*, vol. 96, pp. 3392–3400, 1994.

[26] C. M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "Emotion recognition based on phoneme classes," *Proceedings of ICSLP, Jeju, Korea*, 2004.

[27] C. M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 293–303, 2005.

[28] J. Hansen and B. Womack, "Feature analysis and neural network-based classification of speech under stress," *IEEE Trans. Speech Audio Process*, vol. 4, pp. 307–313, 1996.

[29] Y.Wang and L. Guan, "An investigation of speech based human emotion recognition," *IEEE 6th Workshop on Multimedia Signal Processing*, 2004.

[30] B. Womack and J. Hansen, "Classification of speech under stress using target driven features," *Elsevier Speech Communication*, vol. 20, pp. 131–150, 1996.

[31] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book*, 2002.

[32] M. Airas, "Tkk aparat: An environment for voice inverse filtering and parameterization," *Logopedics Phoniatrics Vocology*, vol. 33, pp. 49–64, 2008.

[33] T. E. Tremain, "The government standard linear predictive coding algorithm: Lpc-10," *Speech Technology Magazine*, pp. 40–49, 1982.

[34] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. Signal Processing*, vol. 39, pp. 411–423, 1991.

[35] M. Airas, H. Pulakka, T. Backstrom, and P. Alku, "A toolkit for voice inverse filtering and parametrisation," *Proc. Interspeech, Lisbon Portugal*, pp. 2145–2148, 2005.

[36] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, pp. 559–601, 1980.

[37] H. M. Teager and S. M. Teager, "A phenomenological model for vowel production in the vocal tract," *Speech Sciences: Recent Advances*, pp. 73–109, 1985.

[38] J. F. Kaiser, "Some useful properties of the teager's energy operators," *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing, (Mineapolis, MN)*, vol. 3, pp. 149–152, 1993.

[39] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On separating amplitude from frequency modulations using energy operators," *Proc. Internat. Conf. on Avoustics, Speech, and Signal Processing (San Francisco, CA)*, 1992.

[40] ——, "Speech nonlinearities, modulations and energy operators," *Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing, (Toronto, Ontario)*, 1991.

[41] ——, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3024–3051, 1993.

[42] ——, "Detecting nonlinearities in speech using an energy operator," *IEEE DSP Workshop, (New Paltz, NY)*, 1990.

[43] J. Hansen, L. G. Ceballos, and J. Kaiser, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment," *IEEE Transactions on biomedical engineering*, vol. 45, no. 3, 1998.

[44] A. Potamianos and P. Maragos, "Speech analysis and synthesis using an am-fm modulation model," *Speech Communication*, vol. 28, pp. 195–209, 1999.

[45] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust am-fm features for speech recognition," *IEEE Signal Processing Letters*, vol. 12, No.9, 2005.

[46] ——, "Modulation features for speech recognition," *IEEE 2002*, 2002.

[47] Y. Chen and C. Lin, "Combining svms with various feature selection strategies," *NIPS 2003, Feature Selection Challenge*, 2003.

[48] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[49] C. Bishop, *Pattern Recognition and Machine Learning*, 2006.

[50] A. A. D. Souza, "Using em to estimate a probability density with a mixture of gaussians."

[51] V. Vapnik, *The nature of statistical learning theory.* Springer-Verlag NY, 1995.

[52] t. f. e. Wikipedia, "Support vector machines," 2012.

[53] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[54] F. Burkhardt, A. Paeschke, M. Rolfes, M. Sedlmeier, and B. Weiss, "A database of german emotional speech," *Interspeech, Lisbon, Portugal*, pp. 1517–1520, 2005.

[55] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," *Proc. Int. Conf. Acoustics Speech Signal Process.*, vol. 4, pp. 941–944, 2007.