ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ, ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΟΡΑΣΗΣ ΥΠΟΛΟΓΙΣΤΩΝ, ΕΠΙΚΟΙΝΩΝΙΑΣ ΛΟΓΟΥ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑΣ ΣΗΜΑΤΩΝ

Αναγνώριση Συναισθήματος μέσω της Φωνής

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Θεοδώρας Χάσπαρη

Επιβλέπων: Πέτρος Α. Μαραγκός
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2010

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ, ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΟΡΑΣΗΣ ΥΠΟΛΟΓΙΣΤΩΝ, ΕΠΙΚΟΙΝΩΝΙΑΣ ΛΟΓΟΥ ΚΑΙ ΕΠΕΞΕΡΓΑΣΙΑΣ ΣΗΜΑΤΩΝ

Αναγνώριση Συναισθήματος μέσω της Φωνής

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Θεοδώρας Χάσπαρη

Επιβλέπων: Πέτρος Α. Μαραγκός
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή στις 20 Ιουλίου 2010.

.....	.....	.....
Πέτρος Μαραγκός Καθηγητής Ε.Μ.Π.	Αθανάσιος Πρωτόπαπας Καθηγητής Ε.Κ.Π.Α.	Γεράσιμος Ποταμιάνος Διευθυντής Ερευνών ΕΚΕΦΕ Δημόκριτος

Αθήνα, Ιούλιος 2010.

.....

Θεοδώρα Χάσπαρη
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Abstract

Emotion Recognition is a part of affective computing, which focuses on facilitating communication between humans and computers. In this diploma thesis, we examine emotion recognition based on speech.

More specifically, basic features of emotion recognition, such as pitch, formants and utterance duration, are studied analytically. Besides these, it is supported that AM-FM modulation features can distinguish the fine variations of emotions in speech. Instant amplitude and frequency are computed through the Energy Separation Algorithm (ESA) based on the Teager Energy Operator (TEO). Statistical moments of them are used as features. These features are strongly smoothed with median filtering in order to remove the redundant information and keep only the essential information needed for emotion recognition.

Experiments are conducted in two databases: the Berlin Database and the Aiginiteio Hospital Database of Emotional Speech, which include seven and five classes of basic emotions respectively. Classification is done with K-means algorithm, GMMs based on expectation maximization and dynamically modified GMMs. Results vary from 30% to 90%. The most powerful features, which produce the best results, seem to be the TEO-Autocorrelation-Envelope, the Area of Instant Amplitude and the Weighted Mean of Instant Frequency.

Keywords: Emotion Recognition, Pitch, Formants, AM-FM demodulation features, GMMs

Ευχαριστίες

Η ενασχόλησή μου με την επεξεργασία φωνής ξεκίνησε στα πλαίσια του μαθήματος Ψηφιακή Επεξεργασία Σημάτων, που διδάσκει ο καθηγητής του Εθνικού Μετσοβίου Πολυτεχνείου Πέτρος Μαραγκός. Η μεταδοτικότητα του καθηγητή μου και το ενδιαφέρον του αντικειμένου με ώθησαν να το διαλέξω ως γενικότερο θέμα της διπλωματικής μου εργασίας. Στο πλάισιο αυτό, μου δόθηκε η ευκαιρία να εργαστώ στο εργαστήριο της Ελληνικής Εταιρείας για την Προαγωγή τής Ψυχιατρικής και των Συναφών Επιστημών (ΕΛ.Ε.Π.Ψ.Υ.Σ.ΕΠ.) στο Αιγινήτειο Νοσοκομείο, υπό την επίβλεψη του καθηγητή της Ιατρικής Κωνσταντίνου Σολδάτου.

Το θέμα της εργασίας αυτής, που είναι η αναγνώριση συναισθήματος από τη φωνή, προέκυψε από τη συνένωση σε ερευνητικό επίπεδο των τομέων του της επιστήμης του Ηλεκτρολόγου Μηχανικού και της Ψυχιατρικής. Η καθοδήγηση του καθηγητή Πέτρου Μαραγκού καθόλη τη διάρκεια της διπλωματικής εργασίας ήταν συνεχής και πολύτιμη. Είναι εξαιρετικός δάσκαλος και θα ήθελα να τον ευχαριστήσω από τα βάθη της καρδιάς μου για τις κατευθύνσεις που μου έδωσε και για όλη τη βοήθειά του όχι μόνο στην πορεία αυτής της μελέτης αλλά και για τις μελλοντικές επιλογές μου.

Επίσης, θα ήθελα να ευχαριστήσω τον καθηγητή Κωνσταντίνο Σολδάτο για την υποστήριξή του στο εργαστήριο της ΕΛ.Ε.Π.Ψ.Υ.Σ.ΕΠ., καθώς και για τις πολύτιμες παρατηρήσεις του στον καθορισμό και τη διεξαγωγή της πειραματικής διαδικασίας.

Στα πλαίσια της εργασίας μου στην ΕΛ.Ε.Π.Ψ.Υ.Σ.ΕΠ. πολύτιμη ήταν η καθοδήγηση του καθηγητή του ΕΚΠΑ Αθανάσιου Πρωτόπαπα στην οργάνωση του εργαστηριακού εξοπλισμού. Καθόλη τη διάρκεια της έρευνας, οι υποδείξεις του ήταν καθοριστικές για την τελειοποίηση της ποιότητας της ηχητικής καταγραφής. Τον ευχαριστώ για αυτό καθως και για τη συμμετοχή του στην εξεταστική επιτροπή.

Επίσης, ευχαριστώ πολύ τον ερευνητή του Δημόκριτου κύριο Γεράσιμο Ποταμιάνο για τη συμμετοχή του στην τριμελή επιτροπή.

Κατά την εκπόνηση της διπλωματικής εργασίας, είχα την ευκαιρία να συνεργαστώ με το Δημήτρη Δημητριάδη, που μου παραχώρησε και ένα κομμάτι κώδικα στο Matlab. Τον ευχαριστώ για την υποστήριξη και τις παρατηρήσεις του, που ήταν χρήσιμες και καίριες. Επιπλέον, θα ήθελα να ευχαριστήσω θερμά το Σταύρο Θεοδωράκη, που με βοήθησε πάρα πολύ σε απορίες και τεχνικά ζητήματα, και όλα τα υπόλοιπα μέλη του εργαστηρίου Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων του ΕΜΠ.

Ουσιαστική ήταν η συμβολή του Νίκου Ρόζου, μελους του εργαστηρίου της ΕΛ.Ε.Π.Ψ.Υ.Σ.ΕΠ., καθώς και των μελών της γραμματείας, Εύης Σπανού και Μαρίας Μίντουλα. Τους ευχαριστώ πάρα πολύ για τη βοήθειά τους.

Πολύτιμη ήταν και η συμμετοχή των παιδιών στις καταγραφές στην ΕΛ.Ε.Π.Ψ.Υ.Σ.ΕΠ., χάρη στους οποίους έγινε η συλλογή της βάσης δεδομένων και θα ήθελα θερμά να τους ευχαριστήσω για το χρόνο που μου αφιέρωσαν: Ε. Αρβανίτη, Ν. Βάζου, Γ. Γκιοξάρη, Θ.

Γουλεάκης, Ε. Δημογεροντάκης, Α. Θεοδωρίδης, Σ. Καλλίνωσης, Ε. Κοντοπόδης, Ε. Κοντός, Χ. Μιναρεντζής, Ι. Παπακωνσταντίνου, Ε. Πολυμενέας, Ρ. Ράιδου, Δ. Ριζάδης, Γ. Σαντίκος, Σ. Στασινόπουλος, Λ. Σταυρίδου, Α. Στρατηγίου, Μ. Τσατσανίφου, Η. Ψαρουδάκης.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

Η αναγνώριση συναισθήματος έχει αναπτυχθεί τα τελευταία χρόνια και αποτελεί μεγάλο μέρος της έρευνας με στόχο την καλύτερη επικοινωνία μεταξύ ανθρώπου και μηχανής. Μεταξύ των ανθρώπων υπάρχουν πολλοί δίαυλοι επικοινωνίας: το περιεχόμενο του λόγου, τα νεύματα, οι κινήσεις προσώπου και σώματος και τα συναισθήματα. Για να γίνει πιο άνετη και σαφής η επικοινωνία ανθρώπου και υπολογιστή, θα μπορούσε ο υπολογιστής να καταλαβαίνει και να αντιδρά κατάλληλα στα συναισθήματα του ανθρώπου. Έχει μάλιστα υποστηριχτεί ότι η συναισθηματική ευφυία στους υπολογιστές είναι πιο σημαντική από την υπολογιστική και τη λεκτική, ώστε να γίνουν οι εφαρμογές πιο φιλικές προς τους ανθρώπους. Η συναισθηματική ευφυία είναι απαραίτητη ώστε να διαπιστωθούν οι προτιμήσεις των ανθρώπων και να προσαρμοστούν οι υπολογιστές στα χαρακτηριστικά κάθε ανθρώπου ξεχωριστά.

Στη συνέχεια, παραθέτουμε ορισμένες εφαρμογές της αναγνώρισης συναισθήματος, όπως αυτές έχουν αναφερθεί στις υπάρχουσες έρευνες.

- Στην ψυχολογία η κατανόηση της επίδρασης του συναισθήματος στη φωνή θα μπορούσε να βοηθήσει το έργο των γιατρών να καταλάβουν καλύτερα τη φύση των συναισθημάτων. Επίσης, ο αυτόματος εντοπισμός συναισθηματικά φορτισμένων στιγμών θα διευκόλυνε τους ψυχολόγους στο έργο τους [54]. Η μελέτη της επίδρασης ψυχολογικών ασταθειών, όπως η κατάθλιψη, στη φωνή θα μπορούσε να ενισχύσει τη διάγνωση των γιατρών [29, 52]. Μία ωραία εφαρμογή για παράδειγμα θα ήταν να μελετάται η φωνή του ασθενούς πριν και μετά από φαρμακευτική αγωγή.

- Στα αυτόματα τηλεφωνικά κέντρα πολλοί άνθρωποι δυσκολεύονται να συννενοηθούν με το μηχάνημα αναγνώρισης φωνής και χάνουν την υπομονή τους, με αποτέλεσμα να μην εκπληρώνεται το αίτημά τους. Αν υπήρχε αυτόματη αναγνώριση του θυμού, της απογοήτευσης και της απαρέσκειας, ο πελάτης θα μπορούσε να οδηγηθεί σε ανθρώπινο αντιπρόσωπο χωρίς να ταλαιπωρηθεί [12, 27, 47]. Παρόμοια εφαρμογή θα μπορούσε να υπάρχει και στους προσωπικούς υπολογιστές και σε άλλα μηχανήματα που αλληλεπιδρούν με τον άνθρωπο, όπως τα ATMs.

- Έχει διαπιστωθεί ότι πολλά αυτοκινητιστικά ατυχήματα έχουν συμβεί εξαιτίας της ασταθούς ψυχολογικής κατάστασης του οδηγού και της δυσκολίας του να διαχειριστεί το θυμό του. Για την ενίσχυση της ασφάλειας κατά την οδήγηση, ένα έξυπνο σύστημα θα μπορούσε να αναγνωρίζει το θυμό, την απογοήτευση, το φόβο και το άγχος στη φωνή του οδηγού και να αντιδράει με το κατάλληλο μήνυμα, ώστε να τον κάνει να συνειδητοποιήσει την ψυχολογική του κατάσταση [28, 49, 75].

- Στον τομέα των τηλεπικοινωνιών, θα μπορούσε να βελτιωθεί η απόδοση των τηλεφώνων στη μετάδοση μη λεκτικής πληροφορίας. Αντί να μεταφέρεται ολόκληρη η εικόνα ενός ανθρώπου, το τηλέφωνο του δέκτη μπορεί να αναγνωρίζει το συναίσθημα του συνομιλητή του και να μεταβάλλει ανάλογα την έκφραση του προσώπου του. Έτσι, υπάρχει μεγάλο κέρδος στο χρόνο και το κοστος μεταφοράς του σήματος και για τα βιντεοτηλέφωνα μπορούν να χρησιμοποηθούν οι υπάρχουσες υποδομές 2ης γενιάς (2G MOBILE PHONES). Η εφαρμογή αυτή είναι γνωστή και ως Voice Driven Emotion Recognizer Mobile-phone (VDERM) [66].

- Μία ακόμη εφαρμογή της αναγνώρισης συναισθήματος είναι σε διαδραστικές ταινίες [56], όπου οι χαρακτήρες του υπολογιστή αλληλεπιδρούν με το χρήστη, που είναι και ο πρωταγωνιστής της ταινίας.

- Η αναγνώριση συναισθήματος θα μπορούσε να βοηθήσει και τους ηθοποιούς για να δοκιμάσουν την υποκριτική τους συνέπεια στα διάφορα συναισθήματα [85].

Στα πλαίσια της διπλωματικής αυτής, μελετάμε χαρακτηριστικά της ανθρώπινης φωνής που μπορούν να διακρίνουν τη συναισθηματική κατάσταση του ομιλητή. Αρχίζουμε από κλασσικά χαρακτηριστικά, όπως το pitch και τα formants και συνεχίζουμε με AM-FM χαρακτηριστικά διαμόρφωσης, όπως η Teager ενέργεια, το στιγμιαίο πλάτος και η στιγμιαία συχνότητα. Στα μεγέθη αυτά υπολογίζονται οι στατιστικές ροπές και οι παράγωγοί τους και εφαρμόζονται μέθοδοι ομαλοποίησης για την εξαγωγή χρήσιμης πληροφορίας συναισθήματος. Τέλος, δοκιμάζουμε την ταξινόμηση προτάσεων που έχουν εκφραστεί με διάφορα συναισθήματα, όπως θυμός, χαρά, λύπη, φόβος, απαρέσκεια και πλήξη με K-means, GMMs και αναδιαμορφούμενα GMMs. Πιο συγκεκριμένα, μελετήθηκαν τα εξής σημεία:

- Έγινε μία επισκόπηση των βασικών χαρακτηριστικών που χρησιμοποιήθηκαν σε προηγούμενες έρευνες, όπως το pitch, τα formants και η διάρκεια εκφοράς και παρατηρήθηκαν διαφορές μεταξύ ορισμένων συναισθημάτων. Βρέθηκε ότι ο θυμός και η χαρά έχουν υψηλές τιμές pitch, ενώ η πλήξη, η λύπη και το ουδέτερο πιο μικρές τιμές. Επίσης, η απαρέσκεια και η λύπη εκφράζονται με πιο αργό λόγο, ενώ η χαρά, ο φόβος και το ουδέτερο και πιο κοφτό λόγο.

- Χρησιμοποιήθηκαν τα χαρακτηριστικά διαμόρφωσης AM-FM για την πιο αποδοτική αναγνώριση συναισθήματος. Τα χαρακτηριστικά αυτά εντοπίζουν ακριβέστερα τις βασικές συχνότητες, τη θεμελιώδη συχνότητα καθώς και τις μεταβολές πλάτους του σήματος φωνής. Πιο αποδοτικά ήταν το εμβαδόν του στιγμιαίου πλάτους (Ampl Area), το εμβαδόν της αυτοσυσχέτισης του στιγμιαίου πλάτους (TEO-Auto-Env) καθώς και ο απλός και σταθμισμένος μέσος όρος στιγμιαίας συχνότητας.

- Στην ταξινόμηση με K-means βρέθηκε ότι υπερτερούν τα ομαλοποιημένα χαρακτηριστικά διαμόρφωσης υπολογισμένα σε 20 κανάλια. Η ομαλοποίηση είναι πιθανό να απομακρύνει την περιττή πληροφορία και να διατηρεί μόνο την πληροφορία που είναι απαραίτητη για την αναγνώριση του συναισθήματος.

- Στην ταξινόμηση με GMMs ο πιο αποδοτικός αριθμός γκαουσιανών ανά κλάση παρατηρήθηκε ότι είναι το 24 και τα πιο ισχυρά χαρακτηριστικά είναι αυτά που έχουν

υπολογιστεί σε 12 κανάλια. Πιστεύουμε ότι ο αριθμός των 12 καναλιών είναι ο ιδανικός ώστε να διατηρείται η χρήσιμη πληροφορία και να μην υπεισέρχεται μεγάλη λεπτομέρεια στους υπολογισμούς.

- Βρέθηκε ότι τα αναδιαμορφούμενα GMMs έχουν τα καλύτερα αποτελέσματα, και αυτό γιατί σε αυτά το πλήθος των ομάδων καθορίζεται δυναμικά. Δεν είναι απαραίτητο το πλήθος των κλάσεων να ισούται με το πλήθος των συναισθημάτων υπό ταξινόμηση, οπότε διανύσματα που αφορούν συναισθήματος μπορεί να ανήκουν σε πολλές ομάδες.
- Κατασκευάστηκε η βάση δεδομένων συναισθηματικής ομιλίας του Αιγινήτειου Νοσοκομείου με 556 προτάσεις, περίπου 111 για 5 βασικά συναισθήματα. Συγκρίθηκε η βάση αυτή με την υπάρχουσα βάση δεδομένων Berlin Database of Emotional Speech.

Κεφάλαιο 2

Το Συναίσθημα από την Πλευρά της Ψυχολογίας

Δύο είναι οι βασικοί δίαυλοι της ανθρώπινης επικοινωνίας [15]: στον πρώτο μεταδίδονται σαφή και στο δεύτερο υποννοούμενα μηνύματα που αντικατοπτρίζουν τις σκέψεις, προθέσεις, ενέργειες και καταστάσεις των ανθρώπων κάθε στιγμή. Εξίσου απο την πλευρά της γλωσσολογίας και της τεχνολογίας έχουν γίνει πολλές προσπάθειες για την αποκωδικοποίηση του πρώτου καναλιού. Για την αποσαφήνιση όμως του δεύτερου, στο οποίο συμπεριλαμβάνεται και η κατανόηση των συναισθημάτων, συνεχίζονται πολλές μελέτες.

Τα συναισθήματα παίζουν σημαντικό ρόλο στους ανωτερους μηχανισμούς, καθώς φαίνεται να καθορίζουν την αντίδρασή τους σε εξωτερικά (συνήθως κοινωνικά) και εσωτερικά ερεθίσματα, που είναι σημαντικά για τις ανάγκες και τους στόχους τους. Σύμφωνα με τον R. Plutchik [64], τα συναισθήματα είναι μία σύνθετη αλυσίδα χαλαρά συνδεδεμένων γεγονότων, που ξεκινάει από ένα ερέθισμα και συνεχίζει με ψυχολογικές μεταβολές, τάσεις για κάποια αντίδραση στο ερέθισμα και τέλος συγκεκριμένη ενέργεια που εξυπηρετεί κάποιο στόχο. Δηλαδή τα συναισθήματα δε συμβαίνουν ποτέ μόνα τους, αλλά αποτελούν αντιδράσεις σε καταστάσεις της ζωής και συνήθως προκαλούν και αντιδράσεις. Ο P. Ekman [24] υποστηρίζει μάλιστα ότι η αντίδραση των ανθρώπων στα συναισθήματα προσαρμόζεται σύμφωνα με το παρελθόν, είτε αυτό αφορά την προγενέστερη ιστορία του ανθρώπινου είδους, είτε τα βιώματα του ίδιου του ατόμου. Ο K.R. Scherer [70] συμπληρώνει ότι τα συναισθήματα είναι επεισόδια αλληλένδετων, συγχρονισμένων αλλαγών στην κατάσταση όλων ή των περισσότερων από τα 5 βασικά οργανικά υποσυστήματα, ως αντίδραση στην αξιολόγηση ενός εξωτερικού ή εσωτερικού ερεθίσματος.

Η βασική ιδιότητα των συναισθημάτων είναι ότι η έκφρασή τους αντικατοπτρίζει μεγάλο όγκο πληροφορίας [24]: το εσωτερικό του ατόμου (σχέδια, αναμνήσεις, κλπ), το πιθανό ερέθισμα στην έκφραση του συναισθήματος αυτού και την πιθανή ακόλουθη αντίδραση (άμεσες συνέπειες, προσπάθειες διακανονισμού, κλπ). Τα συναισθήματα είναι απαραίτητα στην απρόσκοπτη ανάπτυξη και ρύθμιση των διαπροσωπικών σχέσεων. Έχει δειχθεί ότι άνθρωποι που πάσχουν από το σύνδρομο Mobius, που αφορά την παράλυση των μυών του προσώπου, αντιμετωπίζουν μεγάλες δυσκολίες στο να αναπτύξουν και να διατηρήσουν κοινωνικές σχέσεις. Το ίδιο έχει βρεθεί και σε ασθενείς που δεν μπορούν να εκφράσουν την προσωδία στο λόγο τους. Αυτό βέβαια δε σημαίνει ότι κάθε φορά που υπάρχει ένα συναίσθημα, εμφανίζεται και το ανάλογο σήμα, καθώς μπορεί να εμποδίζεται η ανάπτυξη αυτού. Παρόλα αυτά υποστηρίζεται [24] ότι αν μπορούσαμε να μετρήσουμε τις περιοχές του εγκεφάλου που

στέλνουν πληροφορία για την έκφραση του προσώπου, θα υπήρχε κάποια δραστηριότητα, ακόμα και όταν το συναίσθημα δε γίνεται εμφανές.

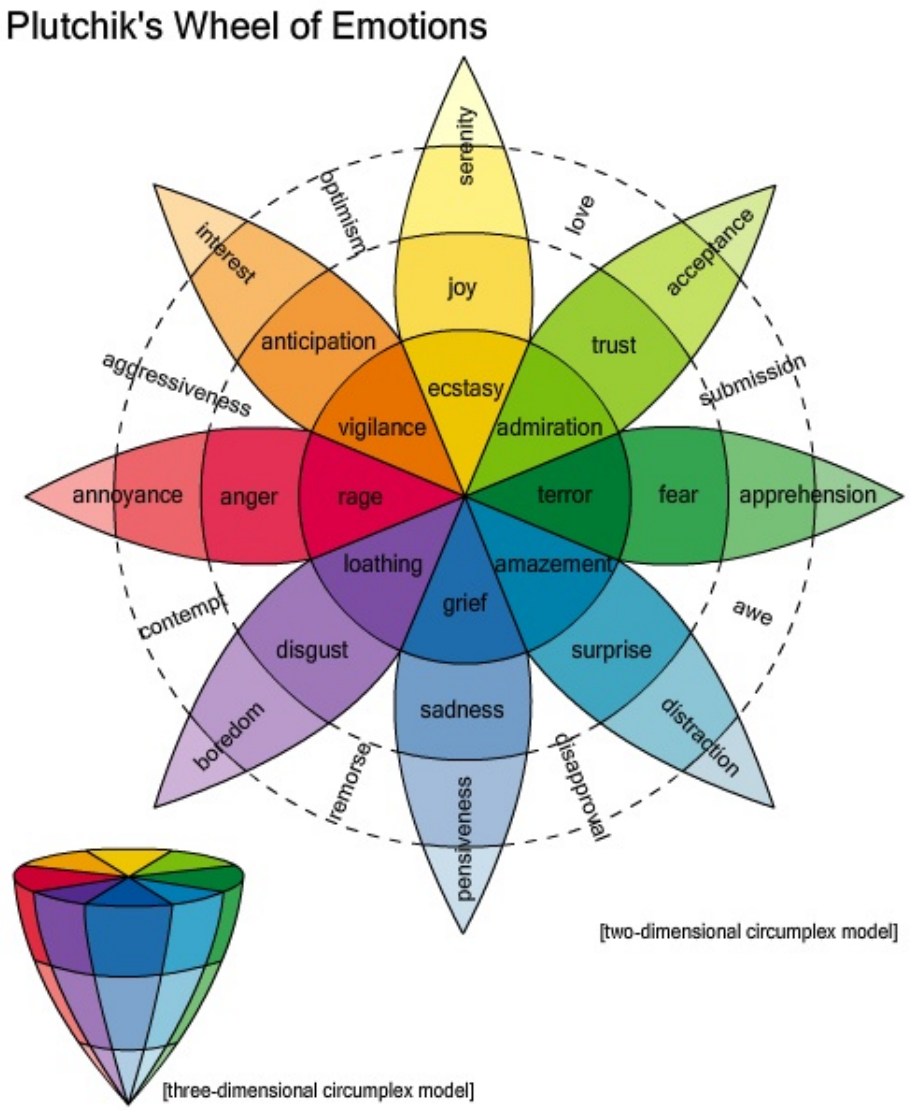
Στο χώρο της ψυχολογίας επικρατούν 2 κυρίαρχοι τρόποι περιγραφής των συναισθημάτων: η πρώτη θεωρία υποστηρίζει ότι υπάρχουν κάποια συναισθήματα που είναι διακριτά, βασικά και παγκόσμια, ενώ η δεύτερη θεωρεί ότι τα συναισθήματα μπορεί να γίνουν αντιληπτά ως σημεία επάνω σε 2 ή 3 διπολικούς άξονες, οπότε και είναι αλληλένδετα [55].

2.1 Βασικά Συναισθήματα

Οι ερευνητές που υποστηρίζουν τη θεωρία αυτή θεωρουν ότι υπάρχουν ορισμένα βασικά, διακριτά και παγκόσμια συναισθήματα, που το καθένα έχει μοναδική φυσιολογική διέγερση, έκφραση, τρόπο διοργάνωσης των γνώσεων και της αντίληψης και κινητοποίηση το οργανισμού. Η διαφοροποίηση μεταξυ τους έγκειται σε φυλογενετικά πρότυπα προσαρμογής και σε αναπτυξιακά νευρικά κυκλώματα. Η κύρια υποστήριξη της θεωρίας αυτής είναι ότι υπάρχουν ορισμένες συνήθεις παρατηρήσεις συναισθημάτων, που φαίνεται να επικρατούν σε όλους τους πολιτισμούς και σε ορισμένα ανώτερα ζώα. Επίσης υπάρχουν συναισθήματα που σχετίζονται και αναγνωρίζονται σε παγκόσμιο επίπεδο με συγκεκριμένες εκφράσεις του προσώπου και εξυπηρετούν συγκεκριμέβες βιολογικές ανάγκες [58]. Κατά καιρούς έχουν διαμορφωθεί διάφορες ομάδες βασικών συναισθημάτων από ερευνητές, όπως φαίνεται στον πίνακα 2.1, σύμφωνα με τους A. Ortony και T.J Turner, που έχουν συγκεντρώσει τις πιο βασικές από τις θεωρίες αυτές.

Πολλοί ψυχολόγοι θεωρούν ότι τα βασικά συναισθήματα μπορούν να συνδυαστούν ή να αναμειχθούν και με τον τρόπο αυτο εξηγείται η μεγάλη ποικιλία των λεκτικών περιγραφών των συναισθημάτων στις διάφορες γλώσσες [71]. Σύμφωνα με τον R. Plutchik [63] όλα τα εμφανιζόμενα συναισθήματα μπορούν να προκύψουν ως συνδυασμοί ή παράγωγα των βασικών. Ο ίδιος μάλιστα δημιούργησε τον 'τροχό των συναισθημάτων', όπως φαίνεται στο σχήμα 2.2, στον οποίο υπάρχουν 8 βασικά διπολικά συναισθήματα. Όπως τα χρώματα, τα βασικά συναισθήματα μπορούν να εκφραστούν σε διαφορετική ένταση και μπορούν να αναμειχθούν ώστε να προκύψουν άλλα συναισθήματα, τα οποία φαίνονται στις ενδιάμεσες θέσεις στην εξωτερική ακτίνα του τροχού.

Το πρόβλημα με την προσέγγιση αυτή είναι ότι ενώ προτείνεται πως τα βασικά συναισθήματα έχουν διαφορετική ψυχολογική βάση, υποστηρίζεται πως αυτά μπορεί να συγχωνευθούν ώστε να προκύψουν παράγωγα συναισθήματα. Επίσης, δεν αναλύονται οι γενικές αρχές του συνδυασμού αυτού, ούτε και τα είδη των μηχανισμών που μπορεί να συμβάλλουν στο συνδυασμό. Το κύριο ομως μειονέκτημα είναι η υποκειμενικότητα στο τι θεωρείται βασικό συναίσθημα [58]. Σύμφωνα με τους A. Ortony και T.J Turner ένα συναίσθημα A μπορεί να θεωρηθεί πιο βασικό από ένα συναίσθημα B, αν το A περιέχει ένα υποσύνολο των ιδιοτήτων του B. Έτσι, αν προσπαθήσουμε να βρούμε τα συστατικά που αποτελούν ένα 'βασικό' συναίσθημα, μπορεί να προκύψει ένα άλλο συναίσθημα με λιγότερα από αυτά τα συστατικά.



Σχήμα 2.1: Ο τροχός των συναισθημάτων του Plutchik [63, 81].

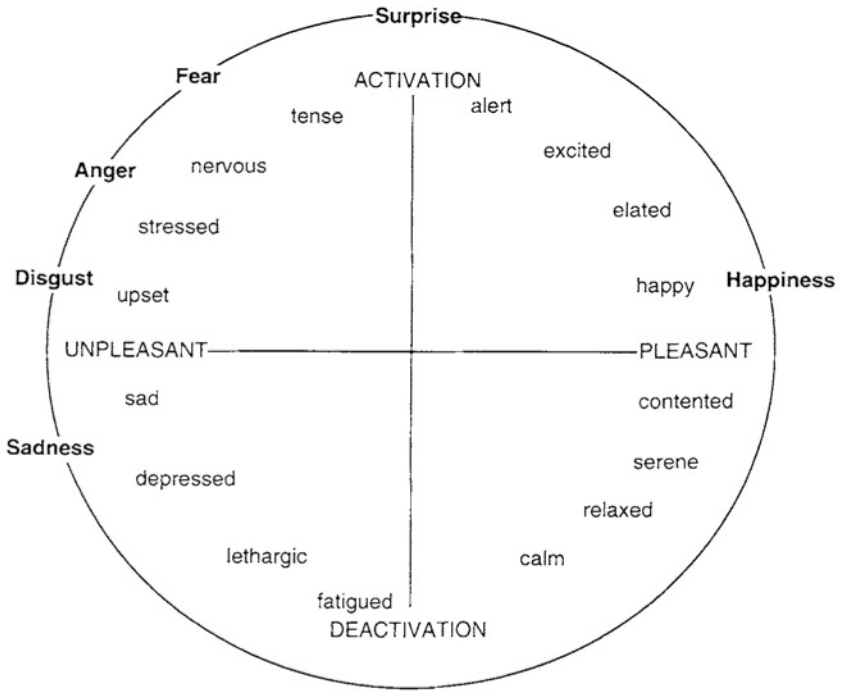
Αναφορά	Βασικό Συναίσθημα	Βάση ταξινόμησης
Arnold(1960)	θυμός, αποστροφή, κουράγιο, απογοήτευση, επιθυμία, απελπισία, φόβος, μίσος, ελπίδα, αγάπη, λύπη	Αντίδραση σε κάθε συναίσθημα
Ekman, Friesen Ellsworth(1982)	θυμός, αηδία, φόβος, χαρά, λύπη, έκπληξη	Παγκόσμιες εκφράσεις προσώπου
Frijda(1986)	επιθυμία, χαρά, ενδιαφέρον, έκπληξη, απορία, θλίψη	Ετοιμότητα αντίδρασης
Gray(1982)	λύσσα και τρόμος, άγχος, χαρά	
Izard(1971)	θυμός, περιφρόνηση, αποστροφή, αγωνία, φόβος, ενοχή, ενδιαφέρον, χαρά, ντροπή, έκπληξη	
James(1884)	φόβος, πένθος, αγάπη, μίσος	Κινήσεις σώματος
McDougall(1926)	θυμός, αποστροφή, αγαλλίαση, φόβος, υποτέλεια, προσφορά, κατάπληξη	Ένστικτα
Mowrer(1960)	πόνος, ευχαρίστηση	Υποσυνείδητο
Oatley, Johnson Laird(1987)	θυμός, αποστροφή, άγχος χαρά, λύπη	
Panksepp(1982)	προσδοκία, φόβος, λύσσα, πανικός	
Plutchik(1980)	αποδοχή, θυμός, προσμονή, αποστροφή χαρά, φόβος, λύπη, έκπληξη	Προσαρμοστικές βιολογικές διαδικασίες
Tomkins(1984)	θυμός, ενδιαφέρον, περιφρόνηση, αποστροφή αγωνία, φόβος, χαρά, ντροπή, έκπληξη	Πυκνότητα νευρωνικής διέγερσης
Watson(1930)	φόβος, χαρά, οργή	
Weiner, Graham(1984)	χαρά, λύπη	

Πίνακας 2.1: Ομάδες βασικών συναισθημάτων από ερευνητές (συγκεντρωτικός πίνακας από τους A. Ortony και T.J Turner)

2.2 Συναισθήματα ως Σημεία σε Διπολικούς Άξονες

Πολλοί από τους πολέμιους της θεωρίας των βασικών συναισθημάτων υποστηρίζουν ότι τα συναισθήματα μπορούν να γίνουν αντιληπτά ως σημεία πάνω σε δύο ή τρεις διπολικούς άξονες. Είναι γεγονός ότι τα βασικά συναισθήματα διαφέρουν ως προς την ένταση, το βαθμό ευχαρίστησης και το βαθμό ενεργοποίησης. Για το λόγο αυτό μπορεί να υποστηριχθεί ότι τα συναισθήματα μπορούν να παρασταθούν σε τρεις διαστάσεις: ευχαρίστηση, διέγερση και δραστικότητα (valence, arousal, potency). Η ευχαρίστηση είναι η κυριότερη διάσταση και πληροφορεί για το αν το συναίσθημα είναι θετικό ή αρνητικό. Η διέγερση αντιπροσωπεύει το βαθμό της ψυχολογικής εμπλοκής του ατόμου στο συναίσθημα αυτό, όπως το βαθμό εγγρήγορσης και συγκίνησης. Τέλος, η δραστικότητα είναι η πιο σπάνια χρησιμοποιούμενη διάσταση και απεικονίζει την αίσθηση ελέγχου του ατόμου στο συναίσθημα.

Το μοντέλο αυτό προτάθηκε αρχικά από το W. Wundt, ο οποίος θεωρούσε ότι οι τρεις αυτές διαστάσεις είναι περιγραφικά χαρακτηριστικά μίας ενιαίας συναισθηματικής κατάστα-



Σχήμα 2.2: Αναπαράσταση των συναισθημάτων σε 2 διαστάσεις: ευχαρίστηση και διέγερση. Ο εσωτερικός κύκλος επεξηγεί τις θέσεις του επιπέδου. Ο εξωτερικός κύκλος τοποθετεί τα πρωτότυπα-βασικά συναισθήματα στις 2 διαστάσεις [67].

σης [2]. Αργότερα ο J.A Russell [67] υποστήριξε την ίδια άποψη, αναφέροντας ότι η έκφραση του προσώπου και η ομιλία για κάθε συναίσθημα μπορεί να κωδικοποιηθεί από την ευχαρίστηση και τη διέγερση. Στο σχήμα 2.2 φαίνεται μία σχηματική απεικόνιση με τους δύο διπολικούς άξονες. Αξιοσημείωτο είναι ότι ορισμένα από τα πρωτότυπα-βασικά συναισθήματα βρίσκονται στα θετικά και αρνητικά άκρα των δύο αξόνων.

Κεφάλαιο 3

Υπάρχουσα Ερευνητική Δραστηριότητα στην Αναγνώριση Συναισθήματος

3.1 Βάσεις Δεδομένων Προφορικής Έκφρασης Συναισθήματος

Οι σημερινές εφαρμογές διεπαφής ανθρώπου και μηχανής έχουν οδηγήσει στην προσπάθεια για την ανάπτυξη αποδοτικών αλγορίθμων για αναγνώριση συναισθήματος καθώς και για τη φυσική έκφραση του συναισθήματος στη σύνθεση φωνής. Για τη μελέτη της επίδρασης του συναισθήματος στη φωνή έχουν δημιουργηθεί πολλές βάσεις δεδομένων, οι οποίες όμως διαφέρουν στην πειραματική διαδικασία συλλογής δεδομένων. Έχουν δημιουργηθεί 4 σημαντικά ερωτήματα κατά τη διάρκεια των ερευνητικών προσπαθειών [21]:

1. Ποιος είναι ο ιδανικός αριθμός μίας βάσης δεδομένων ως προς τα συναισθήματα που περιέχει και τους ανθρώπους που τα εκφράζουν.

2. Ποια πρέπει να είναι η φύση του υλικού: το συναίσθημα μπορεί να είναι αυθόρμητο ή να προσομοιάζεται θεατρικά, επίσης μπορεί να προκαλείται από τον ερευνητή ή από το περιβάλλον.

3. Αρκεί μόνο η φωνητική έκφραση του συναισθήματος ή πρέπει να προστεθούν και άλλα κανάλια (π.χ. εικόνα), ώστε να υπάρχει μεγαλύτερη ακρίβεια.

4. Με ποιο τρόπο θα σημειωθούν και θα περιγραφούν τα συναισθήματα σε μεγάλες βάσεις δεδομένων με υλικό μεγάλης διάρκειας.

Όπως περιγράφεται και σε δημοσιεύσεις [85], τα συναισθήματα που έχουν εξετασθεί στις περισσότερες βάσεις δεδομένων είναι θυμός, λύπη, χαρά, φόβος, απαρέσκεια, έκπληξη και άγχος, που αναλυτικά φαίνονται στον πίνακα 3.1 .

Για τη δημιουργία βάσεων συναισθηματικής ομιλίας χρησιμοποιούνται τρία βασικά είδη λόγου: ο προσποιητός, ο προκλητός και ο φυσικός λόγος. Στον *προσποιητό λόγο* επαγγελματίες ηθοποιοί ή και ερασιτέχνες καλούνται να εκφωνήσουν προκαθορισμένες προτάσεις ή μικρά κείμενα με συγκεκριμένη συναισθηματική κατάσταση. Ο *προκλητός λόγος* βασίζεται

Συναίσθημα	Πλήθος βάσεων Δεδομένων
θυμός	26
λύπη	22
ευτυχία	13
φόβος	13
απαρέσκεια	10
χαρά	9
έκπληξη	6
πλήξη	5
άγχος	3

Πίνακας 3.1: Τα συναισθήματα που έχουν καταγραφεί στις βάσεις δεδομένων [85]

σε φανταστικά σενάρια τα οποία οι εκφωνητές πρέπει να φανταστούν ότι βιώνουν, οπότε και να προκληθούν τα αντίστοιχα συναισθήματα στην ομιλία τους. Τέλος, ο *φυσικός λόγος* προκαλείται από πραγματικές καταστάσεις που συμβαίνουν στον ομιλητή. Στη συνέχεια, αναφέρουμε αντιπροσωπευτικές έρευνες που έχουν γίνει για κάθε είδος ομιλίας. Πλήρης περιγραφή των βάσεων δεδομένων συναισθηματικής ομιλίας έχει δημοσιευτεί σε πολλές έρευνες [21, 85, 84].

3.1.1 Προσποιητός Λόγος: Εκφώνηση Προκαθορισμένων Προτάσεων και Κειμένων

Η εκφώνηση προκαθορισμένων προτάσεων και κειμένων είναι η πιο συνήθης και εύκολα υλοποιήσιμη τεχνική που έχει χρησιμοποιηθεί από πολλούς ερευνητές. Το μεγάλο πλεονέκτημα είναι ότι μπορούν να ελεγχθούν οι συνθήκες καταγραφής και να εξασφαλιστεί κατάλληλος εξοπλισμός, ώστε να μειώνεται ο θόρυβος και να υπάρχει όσο το δυνατόν καλύτερη ποιότητα ήχου.

Ως μία πρώτη τεχνική χρησιμοποιούνται *προτάσεις με ουδέτερο περιεχόμενο*, οι οποίες εκφωνούνται με διαφορετικά συναισθήματα [7]. Σε έρευνα στο Τεχνικό Πανεπιστήμιο στο Βερολίνο εκφωνήθηκαν 10 προτάσεις που χρησιμοποιούνται στην καθημερινότητα με 6 συναισθήματα (θυμός, φόβος, χαρά, λύπη, απαρέσκεια, πλήξη) και το ουδέτερο. Οι ερευνητές αυτοί είναι υπέρ της προσποιητής έκφρασης συναισθημάτων, γιατί είναι ευκολότερα υλοποιήσιμη, και αμεσότερη. Θεωρούν μάλιστα ότι οι ομιλητές πρέπει να εκφράζουν ίδιες φράσεις για όλα τα συναισθήματα, έτσι ώστε να είναι πιο εύκολη η σύγκριση μεταξύ των συναισθημάτων και μεταξύ των ομιλητών, χωρίς να εμπλέκεται το περιεχόμενο μίας φράσης. Το μειονέκτημα σε αυτή την προσέγγιση είναι η δυσκολία των ομιλητών να παράγουν συναισθηματικό λόγο που χαρακτηρίζεται από φυσικότητα, αφού στην πραγματικότητα δε θα βρίσκονται στην κατάλληλη συναισθηματική κατάσταση. Για το λόγο αυτό τους ζητήθηκε να θυμηθούν μία πραγματική εμπειρία από το παρελθόν και να επαναφέρουν το συναίσθημα που τους είχε προκαλέσει. Έτσι, ειπώθηκαν 10 προτάσεις από 10 ανθρώπους (5 άντρες και 5 γυναίκες) με όλα τα μελετούμενα συναισθήματα. Παράδειγμα των προτάσεων αυτών είναι: 'Τα σαββατοκύριακα πηγαίνω πάντα στο πατρικό μου σπίτι και βλέπω τους δικούς μου.', 'Γιατί βάλατε τις τσάντες κάτω από το τραπέζι·', 'Θέλω μόνο να τελειώσω τη δουλειά και μετά να ξεκουραστώ.'. Στη συνέχεια, οι προτάσεις αυτές αξιολογήθηκαν ως προς το

συναίσθημα που απεικονίζουν και ως προς τη φυσικότητά τους και διατηρήθηκαν αυτές για τις οποίες η αναγνώριση του συναισθήματος ήταν πάνω από 80% και η φυσικότητα πάνω από 60%. Η βάση δεδομένων που κατασκευάστηκε με βάση τους παραπάνω κανόνες είναι γνωστή ως Berlin Database of Emotional Speech και έχει χρησιμοποιηθεί σε πολλές έρευνες λόγω της απλότητάς της (περίπου 500 wav αρχεία) και της ελεύθερης πρόσβασης μέσω internet.

Μία άλλη τεχνική είναι η *εκφώνηση προτάσεων με αντιπροσωπευτικό περιεχόμενο για κάθε συναίσθημα* [57]. Στο Εθνικό Πανεπιστήμιο της Σιγκαπούρης μελετήθηκαν 6 συναισθήματα (θυμός, απαρέσκεια, φόβος, χαρά, λύπη, έκπληξη), για καθένα από τα οποία εκφωνήθηκαν 10 προτάσεις από 12 ανθρώπους με 2 διαφορετικές διαλέκτους: Burmese και Mandarin. Παραδείγματα προτάσεων θυμού είναι: 'Έρχεσαι πάντα καθυστερημένος!' και 'Είναι άδικο αυτό!', ενώ απαρέσκειας είναι: 'Δε μου αρέσει το φαί αυτό.', 'Δε θέλω να βγω από το σπίτι.'. Οι δύο διάλεκτοι επιλέχθηκαν γιατί ήταν οι πιο συνηθισμένες στην περιοχή του πανεπιστημίου και στην έρευνα αυτή εκφράζεται η πεποίθηση ότι η γλώσσα δεν επηρεάζει τη συναισθηματική ομιλία. Έχουν προκύψει ερωτήματα [12] και έχουν γίνει έρευνες [69] σχετικά με την πολιτισμική επίδραση στη διαμόρφωση του συναισθήματος. Έχει διαπιστωθεί ότι σε γενικές γραμμές η φωνητική έκφραση συναισθημάτων καθορίζεται από ψυχοβιολογικούς μηχανισμούς, γιατί άνθρωποι που μιλούν διαφορετικές γλώσσες μπορούν να αναγνωρίσουν με μεγάλη ακρίβεια τα συναισθήματα. Υποστηρίζεται παρόλα αυτά ότι οι θεμελιώδεις συχνότητες έκφρασης των φωνημάτων, οι διαμορφώτριες συχνότητες, η άρθρωση, ο τονισμός και ο ρυθμός επηρεάζουν την έκφραση συναισθήματος, που βρίσκει διαφορές ανάμεσα στις γλώσσες.

Σε μία άλλη έρευνα στο Πανεπιστήμιο Delft της Ολλανδίας [12] καταγράφηκε ήχος και εικόνα από 25 ερασιτέχνες για 21 διαφορετικά συναισθήματα, όπως φαίνονται στον πίνακα 3.2. Η καταγραφή έγινε σε έναν ειδικά διαμορφωμένο χώρο, όπου χρησιμοποιήθηκε κάμερα ταχείας καταγραφής και μικρόφωνο πυκνωτή. Το ατόμο καθόταν μπροστά στην κάμερα, αλλά υπήρχε σε γωνία 45^o καθρέφτης με τη βοήθεια του οποίου η ίδια κάμερα κάλυπτε και το προφίλ. Δε χρησιμοποιήθηκαν 2 κάμερες για να αποφευχθούν προβλήματα χρονισμού. Από κάθε άτομο ζητήθηκε να διαβάσει ιστορίες με έντονο συναισθηματικό περιεχόμενο και έπειτα από καθεμία να εκφωνήσει 5 κατάλληλες προτάσεις ως πιθανές αντιδράσεις για τη συγκεκριμένη κατάσταση. Συνολικά μαζεύτηκαν 105 προτάσεις από κάθε άτομο. Οι ερευνητές αναγνωρίζουν ότι η ποιότητα του συνισθηματικού λόγου μπορεί να μην είναι πάντα τόσο γνήσια, παρόλα αυτά υποστηρίζουν ότι στην καθημερινή ζωή είναι πολύ δύσκολο να εντοπιστούν μεμονωμένα συναισθήματα, αφού συνήθως υπάρχει μία μείξη συναισθημάτων στις ανθρώπινες αντιδράσεις.

Τέλος, έχουν γίνει έρευνες όπου παρουσιάζονται *σενάρια στα οποία εκλύονται διάφορα συναισθήματα* και στη συνέχεια πρέπει να εκφωνηθούν κείμενα με προκαθορισμένο περιεχόμενο. Το πλεονέκτημα των κειμένων σε σχέση με τις μεμονωμένες προτάσεις είναι ότι ο εκφωνητής μπορεί να επιτύχει πιο φυσικά το ζητούμενο συναίσθημα. Στο Πανεπιστήμιο της Νότιας Καλιφόρνιας [8], επιλέχθηκαν σενάρια με διαλόγους από θεατρικά έργα και ζητήθηκε από 10 ηθοποιούς σε ζευγάρια ανά 2 να ηχογραφήσουν προκαθορισμένους διαλόγους για 3 συναισθήματα (χαρά, λύπη, θυμός) και το ουδέτερο.

Αρ.	Συναίσθημα	Αρ.	Συναίσθημα	Αρ.	Συναίσθημα
1	θαυμασμός	8	αποστροφή	15	ενόχληση
2	διασκέδαση	9	απαρέσκεια	16	ενδιαφέρον
3	θυμός	10	δυσαρέσκεια	17	ευχάριστη έκπληξη
4	πλήξη	11	γοητεία	18	δυσάρεστη έκπληξη
5	περιφρόνηση	12	φόβος	19	ικανοποίηση
6	επιθυμία	13	οργή	20	λύπη
7	απογοήτευση	14	χαρά	21	έμπνευση

Πίνακας 3.2: Τα συναισθήματα που έχουν καταγραφεί στη βάση δεδομένων του Πανεπιστημίου Delft [12]

3.1.2 Φυσικός Λόγος

Το μειονέκτημα της προσποιητής έκφρασης συναισθήματος είναι ότι δεν έχει βρεθεί η σχέση μεταξύ του θεατρικού λόγου και της αυθόρμητης έκφρασης στην έκλυση συναισθήματος. Για το λόγο αυτό δε γνωρίζουμε κατά πόσο τα αποτελέσματα που προκύπτουν από τις βάσεις δεδομένων που στηρίζονται στη θεατρική ομιλία, μπορούν να είναι αποδοτικά σε καθημερινές εφαρμογές.

Οι βάσεις με φυσικό συναισθηματικό λόγο είναι πιο λίγες και περιορισμένες στην ποικιλία των συναισθημάτων που περιέχουν. Έχει προταθεί η συλλογή δεδομένων από τηλεοπτικές σειρές και έργα, όπως στη βάση Bellfast [21] και EmoTV [1]. Από τα τηλεφωνικά κέντρα έχει καταγραφεί η φωνή των πελατών για τον εντοπισμό θυμού [47]. Επίσης, έχει καταγραφεί η φωνή των πιλότων αεροσκαφών την ώρα της πτήσης για τον εντοπισμό του άγχους στη βάση SUSAS (Speech Under Simulated and Actual Stress) [35].

Η βάση *Belfast natural* [21] περιέχει συζητήσεις ανθρώπων από τηλεοπτικά προγράμματα κυρίως κοινωνικής και θρησκευτικής φύσης. Τα δεδομένα είναι ακουστικά και οπτικά. Η βάση αυτή διέπεται από τον κανόνα οι άνθρωποι να μιλάνε ή έστω να φαίνεται ότι μιλάνε με απόλυτα φυσικό τρόπο και αυτά που λένε πρέπει να τοποθετούνται στα πλαίσια μίας συζήτησης και όχι ενός μονολόγου ή ενός υπαγορευμένου κειμένου. Κύριος στόχος είναι η παρουσιάση των συναισθημάτων όπως προκύπτουν στην καθημερινή ζωή και όχι στην πρωτογενή και υπερβολική μορφή τους (full-blown emotion). Έτσι, μπορούν να προκύψουν ανάμεικτες συναισθηματικές καταστάσεις ή πολλές χρειές ενός συναισθήματος. Αυτή η αβεβαιότητα και η διαβάθμιση των συναισθημάτων έχει αποτυπωθεί στη βάση δεδομένων με ειδικούς περιγραφητές. Η εφαρμογή "Feeltrace" βοήθησε προς την κατεύθυνση αυτή, καθώς ζητήθηκε από ανθρώπους να αξιολογήσουν τα εκλυόμενα συναισθήματα στο συνεχή χώρο δυναμικότητας - εκτίμησης (activation-evaluation) [14]. Μάλιστα βρέθηκε ότι με βάση την απεικόνιση του συναισθήματος στο συνεχή χώρο υπάρχει μεγαλύτερη συμφωνία μεταξύ των ερωτηθέντων απ'οτι στο διακριτό χώρο [21]. Τέλος, εφόσον η βάση Belfast natural είναι οπτικοακουστική, μελετήθηκαν 4 εκδοχές δειγμάτων της: οπτικοακουστική, οπτική, ακουστική, φιλτραρισμένη ακουστική (όπου αποκρύπτεται το περιεχόμενο του λόγου, αλλά δεν επηρεάζεται η προσωδία και η ποιότητα της φωνής) [22]. Η αξιολόγηση του συναισθήματος έγινε πάλι στο συνεχή χώρο με το "Feeltrace" και συγκρίθηκαν οι 3 τελευταίες εκδοχές με την οπτικοακουστική. Ως προς τη δυναμικότητα, παρατηρείται ότι οι ακραίες τιμές της εντοπίζονται σχετικά καλά και στις 3 εκδοχές. Παρόλα αυτά στις ενδιάμεσες τιμές δυναμικότητας το ακουστικό σημα είναι πιο έγκυρο σε σχέση με τα υπόλοιπα δύο. Ως προς την

εκτίμηση, η οπτικοακουστική εκδοχή παρουσιάζει μεγαλύτερη εκτίμηση από τις υπόλοιπες τρεις, ενώ η φιλτραρισμένη ακουστική εμφανίζει πολύ μικρή θετική εκτίμηση.

Η βάση *EmoTV* έχει παρόμοια δομή με τη Belfast natural και δημιουργήθηκε από αποσπάσματα τηλεοπτικών συνεντεύξεων στα γαλλικά. Η περιγραφή των συναισθημάτων στη βάση αυτή έγινε με το ANVIL και βασίστηκε στην απεικόνιση των συναισθημάτων στο διακριτό χώρο (σε 14 κλάσεις) [22]. Στην καθημερινή ζωή δεν είναι πάντα διακριτή η παρουσία συγκεκριμένου συναισθήματος, για αυτό και υιοθετήθηκε μία ακόμη ταξινόμηση στις εξής καταστάσεις: μεμονωμένο συναίσθημα, ταυτόχρονα ανακατεμένα συναισθήματα (προκύπτουν μαζί την ίδια στιγμή), σειριακά ανακατεμένα συναισθήματα (προκύπτουν το ένα μετά το άλλο), αιτιοκρατικά συναισθήματα (το ένα προκαλεί το άλλο). Παρατηρήθηκε ότι μόνο το 46% των δειγμάτων απεικόνιζε ένα μεμονωμένο συναίσθημα, το 33% περιείχε μείγμα ή ακολουθία συναισθημάτων, ενώ για το 21% η αξιολόγηση δεν έδινε κοινό αποτέλεσμα. Επίσης, τα συναισθήματα δεν προκύπτουν πάντα στην πρωτογενή μορφή τους όπως στη θεωρία (αντίδραση σε εξωτερικό ερέθισμα, συγχρονισμός σωματικών κινήσεων, επίδραση στη σκέψη). Έτσι μία δεύτερη ταξινόμηση [19] μπορεί να είναι ως προς την εξέλιξή τους: επεισοδιακό (episodic), υποβόσκον (simmering), φευγαλέο (flitting), διαχεόμενο από τη διάθεση (mood). Το μεγαλύτερο ποσοστό των συναισθημάτων ήταν επεισοδιακά, ακολουθούν τα υποβόσκοντα, τα διαχεόμενα από τη διάθεση και τέλος τα φευγαλέα.

Η εφαρμογή της αναγνώρισης συναισθήματος στα *αυτόματα τηλεφωνικά κέντρα* μπορεί να βελτιώσει την εξυπηρέτηση των πελατών, μεταθέτοντας αυτούς που εκφράζουν αρνητικά συναισθήματα σε ανθρώπινο υπάλληλο. Σε μία έρευνα [47] απομονώθηκαν προτάσεις ανθρώπων που συνομιλούσαν με αυτόματο μηχάνημα σε τηλεφωνικό κέντρο. Το πλεονέκτημα αυτής της τεχνικής συλλογής δεδομένων είναι ότι οι προτάσεις των χρηστών είναι σύντομες και με συγκεκριμένο περιεχόμενο. Έτσι, είναι εύκολη η αξιολόγησή τους και η αναγνώριση συναισθήματος από το περιεχόμενό τους. Μετά την καταγραφή των προτάσεων, ζητήθηκε από ανθρώπους να αξιολογήσουν τα δεδομένα σε δύο κατηγορίες: αρνητική (π.χ. για συναισθήματα θυμού και απογοήτευσης) και μη-αρνητική (π.χ. για συναισθήματα χαράς και ευχαρίστησης). Το αποτέλεσμα έδειξε κάποια ασυμφωνία μεταξύ των ατόμων που συμμετείχαν στην έρευνα, αποδεικνύοντας για μία ακόμα φορά ότι είναι δύσκολο ακόμα και για τους ανθρώπους να αναγνωρίσουν το συναίσθημα στη φωνή. Σε παρόμοια έρευνα [53] επιχειρήθηκε η συλλογή δεδομένων από τηλεφωνικό κέντρο για τη διάκριση ουδέτερου και θυμωμένου λόγου. Στις προτάσεις που απομονώθηκαν, παραμελήθηκαν τα άλλα συναισθήματα και επιλέχθηκαν οι λιγότερο διφορούμενες προτάσεις, όπως αυτές αξιλογήθηκαν από 9 άτομα. Το πλεονέκτημα της τεχνικής αυτής σε σχέση με την καταγραφή εκπομπών από την τηλεόραση είναι ότι οι προτάσεις είναι περισσότερο καθορισμένες και δεν υπάρχει τόσο μεγάλη απαίτηση περιγραφής συναισθημάτων. Παρόλα αυτά η εμβέλεια των βάσεων αυτών είναι πιο περιορισμένη και δεν μπορεί να αντικατοπτρίσει όλα τα συναισθήματα που υπάρχουν στις βάσεις από τις τηλεοπτικές εκπομπές.

Για τη μελετη της επίδρασης του άγχους στη φωνή κατασκευάστηκε η βάση δεδομένων *SUSAS (Speech Under Simulated and Actual Stress)* [34]. Εκτός από την προσομοίωση αργού, γρήγορου και θυμωμένου λόγου, στην έρευνα αυτή επιχειρήθηκε και η καταγραφή φωνής κάτω από πραγματικές συνθήκες άγχους, όπως: α)απαιτητική εργασία στον υπολογιστή β)έλεγχο πτήσης και παρακολούθηση στόχου γ)αγχώδη παιχνίδια σε λούνα-παρκ, όπως το τρενάκι του τρόμου. Επίσης, συλλέχθηκαν κάποιες καταγραφές από τη Σχολή Ψυχιατρικής του Emory Medical University, όπου συμμετείχαν ασθενείς που πάσχουν από κατάθλιψη και στις οποίες απομονώθηκαν προτάσεις αγχώδους λόγου. Η βάση αυτή περιέ-

χει ένα πλούσιο υλικό για την επίδραση του άγχους στη φωνή και έχει χρησιμοποιηθεί σε πολλές έρευνες που ασχολούνται με το θέμα αυτό.

Οι βάσεις συναισθηματικής ομιλίας με φυσικό λόγο έχουν το μεγάλο πλεονέκτημα ότι αντικατοπτρίζουν τις αντιδράσεις των ανθρώπων στην καθημερινή ζωή όσο γίνεται πιο φυσικά, όμως η ποιότητα του ήχου και της εικόνας μπορεί να μην είναι ικανοποητική και μπορεί να υπάρχει θόρυβος στο χώρο καταγραφής. Επίσης, τίθενται ζητήματα προστασίας του ιδιωτικού απορρήτου και της πνευματικής ιδιοκτησίας. Για τους λόγους αυτούς, η δημιουργία τέτοιων βάσεων δεδομένων είναι αρκετά δύσκολη και μπορεί να μην έχει πάντα την επιθυμητή ποιότητα για μετέπειτα επεξεργασία. Όπως υποστηρίζεται και από πολλούς ερευνητές [21], η φυσικότητα στην έκφραση συναισθήματος είναι πολύ σημαντική και δεν μπορεί να εξομοιωθεί με απλές προσομοιώσεις της πραγματικότητας. Παρόλα αυτά, έχουν αναπτυχθεί εξελιγμένες τεχνικές έκλυσης και έκφρασης συναισθήματος, ώστε να δημιουργηθούν έγκυρες πηγές δεδομένων, οι οποίες ξεπερνούν και τα προηγούμενα προβλήματα.

3.1.3 Προκλητός Λόγος

Για να αντιμετωπιστούν τα προβλήματα που αφορούν τις βάσεις δεδομένων φυσικού λόγου, προτάθηκε η πρόκληση συναισθήματος υπό συγκεκριμένες συνθήκες και σε ελεγμένο χώρο ("Mood induction procedure-MIP"). Τα συναισθήματα που εκλύονται παράγονται με φυσικό τρόπο μέσα από διάφορες διαδικασίες, που μπορεί να είναι οι εξής [30, 17]:

- Τεχνικές MIP στις οποίες το συναίσθημα εκλύεται ελεύθερα μέσα από πνευματικές διαδικασίες, όπως φαντασία ή ακόμα και ύπνωση.

- Τεχνικές MIP στις οποίες η καθοδήγηση προς μία συναισθηματική κατάσταση γίνεται με παροτρύνσεις, προτάσεις, αρνητική ή θετική αξιολόγηση.

- Τεχνικές MIP όπου χρησιμοποιείται υλικό κατάλληλο για να προκαλέσει συναισθήματα, όπως μουσικά κομμάτια ή ταινίες.

- Τεχνικές MIP στις οποίες υπάρχει ανάγκη για εκπλήρωση ενός έργου (π.χ. ενός παιχνιδιού στον υπολογιστή ή ενός παιχνιδιού με τουβλάκια), οπότε η επιτυχία ή η αποτυχία εκλύουν και τα ανάλογα συναισθήματα.

- Τεχνικές MIP που χρησιμοποιούν τεχνητά διεγερτικά φάρμακα για να επιφέρουν ψυχολογικές καταστάσεις σχετιζόμενες με συγκεκριμένα συναισθήματα.

Σε μία έρευνα προς αυτή την κατεύθυνση [16, 17], ζητήθηκε από δύο ανθρώπους κάθε φορά σε ένα απομονωμένο δωμάτιο ο καθένας να ανταγωνιστούν παίζοντας ένα παιχνίδι στον υπολογιστή. Βοηθώντας ή βάζοντας εμπόδια στους ανθρώπους αυτούς, μπορεί να γίνει η έκλυση αρνητικών ή θετικών συναισθημάτων. Σημειώνεται ότι οι άνθρωποι αυτοί γνωρίζουν ότι οι αντιδράσεις τους καταγράφονται, αλλά δε γνωρίζουν ότι η έκλυση συναισθήματος είναι ο αυτοσκοπός της έρευνας, αντίθετα αφήνονται να πιστεύουν ότι στόχο αποτελεί ή δοκιμή του παιχνιδιού. Με τον τρόπο αυτό τα συναισθήματα που προκύπτουν είναι φυσικά και αυθόρμητα, αφού οι άνθρωποι δεν υποχρεούνται να τοποθετήσουν τον εαυτό τους σε συγκεκριμένες συναισθηματικές καταστάσεις. Επίσης, επειδή τα ηλεκτρονικά παιχνίδια είναι από μόνα τους εθιστικά στο χρήστη, είναι δυνατή η έκλυση πολλών συναισθημάτων. Η εξωτερική παρέμβαση κόβοντας το παιχνίδι, δίνοντας λανθασμένη πληροφορία για το χρόνο που

απομένει, προσφέροντας ένα βραβείο για το νικητή, μπορεί να παρακινήσει περισσότερο τους ανθρώπους που συμμετέχουν στην έρευνα και να δώσει αξιόπιστα αποτελέσματα. Τέλος, διασφαλίζεται και η ποιότητα του ήχου, αφού το δωμάτιο καταγραφής είναι κατάλληλα μονωμένο, και τα σήματα φωνής των δύο παιχτών δεν εμπλέκονται μεταξύ τους, γιατί παίζουν σε διαφορετικά δωμάτια και έχει ο καθένας άλλο μικρόφωνο.

Παρόμοια τεχνική χρησιμοποιήθηκε και για την αναγνώριση του άγχους από τη φωνή στο πλαίσιο έρευνας για τον εντοπισμό του άγχους κατά τη διάρκεια της οδήγησης [28]. Ζητήθηκε από 4 ανθρώπους να οδηγήσουν σε ένα προσομοιωτή στο Nissan's Cambridge Research Lab, ενώ παράλληλα έπρεπε να κάνουν προσθέσεις αριθμών, των οποίων το άθροισμα είναι μικρότερο από 100 (ώστε να διατηρηθεί σε παρόμοιο επίπεδο η δυσκολία του έργου). Ρυθμίζονταν δύο παράμετροι του πειράματος: η ταχύτητα του οχήματος (60m.p.h, 120m.p.h) και η συχνότητα των ερωτήσεων (1 ερώτηση/ 9sec, 1 ερώτηση/ 4sec). Έτσι δημιουργήθηκαν 4 πειραματικές συνθήκες, μέσω των οποίων συλλέχθηκαν 598 προτάσεις μέσης διάρκειας 1.6sec. Υπήρξε ένα δίλλημα ως προς τον τρόπο ομαδοποίησης των προτάσεων, που μπορούσε να γίνει ανάλογα είτε με τις πειραματικές συνθήκες (cause-type description) είτε με την αξιολόγηση των προτάσεων από τρίτα άτομα (effect-type description). Προτιμήθηκε το πρώτο με σκοπό να ερευνηθεί αν υπάρχει αλγόριθμος που να εκπαιδευτεί ώστε να διακρίνει τα διάφορα επίπεδα άγχους, όπως αυτά προέκυψαν από τη συγκεκριμένη πειραματική διαδικασία.

Ο προκλητός λόγος παραγωγής συναισθήματος μπορεί επίσης να βασιστεί στον αυτοσχεδιασμό. Σε πείραμα στο Πανεπιστήμιο της Νότιας Καλιφόρνιας [8, 10] ζητήθηκε από ανθρώπους να αντιδράσουν σε υποτίθεμενα γεγονότα με συγκεκριμένο συναίσθημα. Οι ομιλητές αντιδρούν σε ζευγάρια ο ένας με τον άλλο και καλούνται να παρομοιάσουν καταστάσεις της καθημερινής ζωής. Για παράδειγμα ο ένας ομιλητής φτάνει στο ταμείο σε μία δημόσια υπηρεσία και συνειδητοποιεί ότι πρέπει να συμπληρώσει μία επιπλέον φόρμα και μετά να ξανακάνει όλη την ουρά, ενώ παράλληλα, ο άλλος ομιλητής απορρίπτει την αίτησή του και προσπαθεί να τον πείσει να ακολουθήσει την ουρά. Έτσι, προκαλούνται συναισθήματα θυμού και απογοήτευσης. Ένα άλλο παράδειγμα είναι μεταξύ δύο φίλων, όπου ο ένας ανακοινώνει στον άλλο ότι παντρεύεται, οπότε απεικονίζεται το συναίσθημα της χαράς. Το πλεονέκτημα της τεχνικής αυτής είναι ότι γίνεται σύνδεση του πειράματος με την καθημερινή ζωή και δημιουργείται κατάλληλο υπόβαθρο κατά την καταγραφή. Από τους ερευνητές θεωρείται επίσης πολύ σημαντική η αλληλεπίδραση μεταξύ δύο ανθρώπων κατά την έκφραση συναισθήματος και πιστεύεται ότι υπερτερεί έναντι του μονολόγου ως προς τη φυσικότητα της έκφρασης. Άλλωστε, όπως έχει υποστηριχτεί από ψυχολόγους [23], κατά τη διάρκεια που οι άνθρωποι έχουν μία έκφραση προσώπου που απεικονίζει ένα συναίσθημα, μπορεί να αρχίσουν να αισθάνονται στην πραγματικότητα το συναίσθημα αυτό. Οι καταγραφές αυτές αξιολογήθηκαν από τρία διαφορετικά άτομα, ώστε να δοκιμαστεί η αξιοπιστία των εκφράσεων για κάθε συναίσθημα. Η αξιολόγηση έγινε με βάση την περογραφή του συναισθήματος και στο συνεχή και στο διακριτό χώρο.

Τέλος, στο Πανεπιστήμιο Aalborg στη Δανία, ομιλητές εκφώνησαν 2 γνωστά κείμενα που αποσκοπούν στην έκλυση συναισθήματος, το κείμενο "North Wind" και το "de Koning" του Godfried Bomans.

3.2 Έρευνες στην Αναγνώριση Συναισθήματος

Η παραγωγή του λόγου απαιτεί ένα συνδυασμό κινήσεων του στόματος, παλμού των φωνητικών χορδών, εισόδου και εξόδου αέρα μέσω του αναπνευστικού συστήματος. Παρόλα αυτά οι άνθρωποι δε λένε τις προτάσεις με τον ίδιο τρόπο κάθε φορά και αυτό γιατί δε βρίσκονται πάντα υπό τις ίδιες συνθήκες. Η πρόσληψη ερεθισμάτων από το περιβάλλον προκαλεί διάφορες συναισθηματικές καταστάσεις, οι οποίες εκφράζονται και μέσω της φωνής. Κατά την επικοινωνία ανθρώπου με άνθρωπο, ο συνομιλητής μπορεί να προσλάβει τα μηνύματα του άλλου και να τα επεξεργαστεί κατάλληλα, ώστε να καταλάβει τη συναισθηματική κατάσταση του συνομιλητή του. Η ίδια επεξεργασία δεν είναι εύκολο να υλοποιηθεί μέσω του υπολογιστή, για αυτό και για την αναγνώριση συναισθήματος μέσω υπολογιστή έχουν διεξαχθεί πολλές έρευνες.

Η αναγνώριση συναισθήματος μέσω της φωνής από τον υπολογιστή μπορεί να εφαρμοσθεί σε αυτόματα τηλεφωνικά κέντρα, όπου ο εντοπισμός θυμού από το χρήστη θα τον οδηγεί σε ανθρώπινο πρόσωπο ώστε να τον εξυπηρετήσει. Επίσης, η ανίχνευση άγχους στη φωνή πιλότων, οδηγών αυτοκινήτων, ελεγκτών εναέριας κυκλοφορίας μπορεί να βοηθήσει στην αποφυγή λαθών. Στις τηλεφωνικές συνδιαλέξεις, από τον εντοπισμό άγχους στη φωνή υπάρχει η δυνατότητα να αποφευχθούν απάτες ή ψεύτικές δηλώσεις. Επίσης, στην ψυχολογία η αναγνώριση συναισθήματος από τη φωνή μπορεί να βοηθήσει τους γιατρούς να κατανήσουν καλύτερα τον τρόπο μηχανισμού παραγωγής συναισθήματος. Μία ακόμη εφαρμογή στην ψυχολογία θα ήταν η απομόνωση συναισθηματικά φορτισμένων προτάσεων μέσα από εκτενείς συνεδρίες, ώστε οι γιατροί να μη χρειάζεται να ψάξουν ολόκληρη την καταγραφή για να βρουν το κομμάτι που τους ενδιαφέρει.

Οι έρευνες με σκοπό την αναγνώριση συναισθήματος εστιάζουν σε ποικίλλα χαρακτηριστικά του λόγου και διαφορετικές μεθόδους ταξινόμησης των συναισθημάτων. Αρχικά θα μελετήσουμε τα χαρακτηριστικά του λόγου που δείχνουν την ύπαρξη συναισθήματος και στη συνέχεια θα δούμε τους τρόπους κατηγοριοποίησης συναισθημάτων που έχουν υλοποιηθεί με βάση τα χαρακτηριστικά αυτά.

3.2.1 Μελέτη Χαρακτηριστικών

Προσωδεία, Χαρακτηριστικά Φωνητικού Σωλήνα και Φωνητικών Χορδών

Τα χαρακτηριστικά της προσωδίας διαμορφώνονται κυρίως από το σχήμα του φωνητικού σωλήνα και σχετίζονται με το ρυθμό και τη μελωδικότητα της φωνής. Η επίδραση του συναισθήματος είναι εμφανής στη φωνή και πιο συγκεκριμένα στη μορφολογία του φωνητικού σωλήνα (vocal tract). Σύμφωνα με πειράματα [36] που αφορούν τους διάφορους τύπους άγχους, κατά τη διάρκεια του ουδέτερου λόγου παρατηρείται μεγαλύτερη ταλάντωση στη φαρυγγική κοιλότητα, ενώ κατά το θυμό η ταλάντωση είναι πιο εμφανής στη γλώσσα και στα χείλια. Τα στοιχεία προσωδίας που έχουν μελετηθεί είναι το pitch, η συχνότητα των formants, η ενέργεια της φωνής και η διάρκεια της ομιλίας.

Pitch

Το pitch είναι η θεμελιώδης συχνότητα του σήματος φωνής που διαμορφώνεται από την ένταση των φωνητικών χορδών, την πίεση και την ταχύτητα του αέρα μεταξύ τους. Πολλές μελέτες φωνής για αναγνώριση συναισθήματος έχουν βασιστεί στο χαρακτηριστικό αυτό, καθώς πιστεύεται ότι είναι πρωταρχικός παράγοντας που επηρεάζεται από το συναίσθημα [90].

Το pitch μελετάται συνήθως σε επίπεδο παραθύρου, παρόλα αυτά έχουν γίνει έρευνες που το μελετούν σε επίπεδο συλλαβής [40], καθώς και σε επίπεδο πρότασης [45].

Σε επίπεδο παραθύρου υπολογίζεται ο μέσος όρος,η μέση τιμή, η τυπική απόκλιση, η μέγιστη και η ελάχιστη τιμή, το εύρος, οι 25% και 75% τιμές του εύρους του pitch [6, 61, 93, 96]. Επίσης, πολλές έρευνες χρησιμοποιούν μία ομαλοποιημένη εκδοχή/περιβάλλουσα του pitch [18, 40, 45, 86, 87, 89] καθώς και την παράγωγό του [18, 86, 87, 89].

Γενικά, οι περισσότερες έρευνες έχουν καταλήξει στο ότι το pitch έχει μεγαλύτερη μέση τιμή και εύρος τιμών στη χαρά και στο θυμό, ενω μικρότερες τιμές στη λύπη και την αποστροφή. Στο θυμό έχουν βρεθεί απότομες διακυμάνσεις του pitch, ενώ στη χαρά το pitch είναι πιο ομαλό. Φθίνουσα πορεία των τιμών του παρατηρείται στα συναισθήματα της λύπης και της αποστροφής. Για το φόβο έχουν παρατηρηθεί πολύ υψηλές τιμές του pitch με μεγάλο εύρος και κανονική διακύμανση [15].

Τέλος, έχει επιχειρηθεί να εξετασθεί η πορεία του pitch κατά τη διάρκεια μιας πρότασης, που βρίκεται με γραμμική παρεμβολή, έτσι ώστε να βρεθεί αν είναι φθίνουσα, σταθερή ή αύξουσα [59]. Βρέθηκε ότι η πορεία του pitch είναι σταθερή στο συναίσθημα της λύπης, σχεδόν σταθερή προς φθίνουσα στην αποστροφή και φθίνουσα στην πλήξη και στο ουδέτερο. Στο θυμό η πορεία του pitch φαίνεται ότι έχει μεγάλες διακυμάνσεις, δηλαδή μπορεί να είναι φθίνουσα, αύξουσα ή και σταθερή. Παρόμοια παρατήρηση αλλά σε μικρότερη έκταση γίνεται για το συναίσθημα της χαράς, ενώ σταθερή ή φθίνουσα πορεία του pitch κατά τη διάρκεια μιας πρότασης φαίνεται να έχει το συναίσθημα του άγχους.

Συχνότητες των Formants

Οι συχνότητες των Formants επηρεάζονται από τη θέση της γλωσσας, των χειλιών και του πίσω μέρους του φωνητικού σωλήνα και έχουν μελετηθεί για αναγνώριση συναισθήματος από πολλές έρευνες [15]. Οι περισσότεροι ερευνητές ασχολούνται με τα δύο πρώτα formants, F1 και F2 [47, 48, 60, 61, 96], ενώ κάποιοι άλλοι λαμβάνουν υπόψη τους και το τρίτο και τέταρτο formant, F3 και F4 [11, 35, 87, 89]. Τα στατιστικά χαρακτηριστικά που υπολογίζονται είναι ο μέσος όρος, η τυπική απόκλιση, το εύρος, το μέγιστο και το ελάχιστο.

Υπό συνθήκες άγχους έχει παρατηρηθεί [35] μεγάλη μεταβλητότητα στα formants. Η επίδραση του άγχους στο λόγο είναι εμφανέστερη με τη χρήση formants για τα φωνήεντα που προφέρονται με τη γλώσσα να βρίσκεται στο μπροστινό μέρος του στόματος, όπως τα 'i', 'y' και σε μικρότερο βαθμό το 'e'. Ο αργός, δυνατός ή αγχώδης λόγος παρουσιάζει τη μεγαλύτερη ολίσθηση στο F1 formant, ενώ το F2 formant αυξάνεται για όλα τα είδη αγχώδους λόγου. Σύμφωνα με άλλη μελέτη που αφορά στα συναισθήματα [96], στη λύπη παρουσιάζεται η μικρότερη μεταβλητότητα για το F1 formant, ενώ για το F2 formant, η μικρότερη μεταβλητότητα υπάρχει στο συναίσθημα της χαράς.

Διάρκεια Ομιλίας

Στις έρευνες για την αναγνώριση συναισθήματος έχουν χρησιμοποιηθεί διάφορες παράμετροι που αφορούν στη διάρκεια της ομιλίας, όπως η διάρκεια των προτάσεων [96], η διάρκεια των συλλαβών [40, 35] και ο ρυθμός ομιλίας [47, 61, 96].

Ο θυμός, η λύπη και η χαρά εκφράζονται συνήθως με προτάσεις μεγαλύτερης διάρκειας σε σχέση με τον ουδέτερο συναίσθημα [96]. Υπολογίζοντας το πηλίκο της διάρκειας σιωπής μεταξύ δύο λέξεων προς τη διάρκεια του λόγου, προκύπτει ότι ο ομιλητής χρησιμοποιεί περισσότερες παύσεις στο συναίσθημα της λύπης σε σύγκριση με τα υπόλοιπα συναισθήματα. Ο ρυθμός ομιλίας μπορεί να υπολογιστεί ως ο αριθμός των φωνηέντων σε ένα δευτερόλεπτο

και σημειώνεται ότι στο θυμό, τη λύπη και τη χαρά υπάρχει μεγαλύτερη μεταβλητότητα του ρυθμού ομιλίας από το ουδέτερο.

Για τα είδη του άγχους [35, 95] παρατηρείται διαφορά στη μέση διάρκεια των λέξεων και των συλλαβών. Πιο συγκεκριμένα, υπολογίζονται οι λόγοι της διάρκειας των συμφώνων προς τη διάρκεια των φωνηέντων, των συμφώνων προς τα ημιφωνήεντα, των φωνηέντων προς τα ημιφωνήεντα. Διαφορές σημειώνονται τόσο στους παραπάνω λόγους, όσο και στη διάρκεια των συλλαβών μέσα σε μία λέξη. Στον ήρεμο και καθαρό λόγο μεγαλύτερη είναι η διάρκεια των φωνηέντων σε σχέση με τη διάρκεια των συμφώνων, ενώ στον πιεσμένο, αγχώδη και θυμωμένο λόγο συμβαίνει το αντίθετο.

Τέλος για το ρυθμό ομιλίας, έχει παρατηρηθεί ότι στο θυμό και στο φόβο είναι σχετικά αυξημένος, ενώ στην αποστροφή πολύ αυξημένος. Στο συναίσθημα της λύπης η ομιλία είναι αργή, ενώ στη χαρά μπορεί να είναι είτε γρήγορη είτε αργή [39].

Ένταση-Ενέργεια Ομιλίας

Η ένταση της ομιλίας φαίνεται να παίζει σημαντικό ρόλο στην ανίχνευση συναισθήματος. Όπως έχει αναφερθεί σε μελέτες [39, 15, 90], στο θυμό και τη χαρά παρατηρείται μεγαλύτερη ένταση της φωνής, ενώ στη λύπη και την αποστροφή η ένταση της φωνής είναι χαμηλότερη. Κανονική ένταση σημειώνεται για το φόβο. Εξίσου σημαντική είναι η ένταση της φωνής για την ανίχνευση των διαφόρων επιπέδων άγχους [35]. Μελετάται η ένταση σε επίπεδο λέξης και σε επίπεδο συλλαβής και βρίσκεται ότι για θυμωμένο και δυνατό λόγο είναι προφανώς αυξημένη, ενώ για καθαρό, ήρεμο και υπό την επήρεια μικρού στρες λόγο είναι μειωμένη.

Για την ενέργεια της ομιλίας στην αναγνώριση συναισθήματος έχουν γίνει πολλές έρευνες [61, 87, 89, 90]. Σε μία μάλιστα [89] έχει βρεθεί οτι χαρακτηριστικά της ενέργειας του σήματος, όπως η μέγιστη τιμή, η διάρκεια του μέγιστου επιπέδου της και η κλίση της, είναι πολύ πιο αποδοτικά σε σχέση με άλλα χαρακτηριστικά του pitch και των formants. Η τελευταία παρατήρηση ισχυει σε πολύ μεγάλο βαθμό για τη γυναικεία φωνή και σε μικρότερο για την αντρική [87].

Σε άλλες έρευνες υπολογίζονται LFPC χαρακτηριστικά, δηλαδή μία λογαριθμική εκδοχή της ενέργειας, που φαίνεται να διαχωρίζουν αρκετά καλά τα συναισθήματα θυμού, έκπληξης, χαράς, φόβου, αποστροφής και λύπης [57].

Γενικά, για τον υπολογισμό της ενέργειας υπάρχουν αρκετές αντιδιαστολές σε σχέση με τη συχνότητα στην οποία διαχωρίζονται αποδοτικά τα συναισθήματα [90]. Ορισμένοι ερευνητές [29] δίνουν μεγάλη σημασία στις χαμηλές συχνότητες για τον υπολογισμό της ενέργειας, ενώ άλλοι [57] πιστεύουν το αντίθετο. Μία πιθανή εξήγηση για αυτό είναι ότι ο αγχώδης ή θυμωμένος λόγος μπορεί να εκφραστεί είτε με μικρή προσπάθεια ταλάντωσης των φωνητικών χορδών, που προκαλεί ομαλή μορφή των formants είτε με τσιριχτή φωνή, πράγμα που μεταθέτει την ενέργεια στις υψηλές συχνότητες.

Παλμός Φωνητικών Χορδών (Glottal Waveform)

Οι φωνητικές χορδές καθορίζουν τον όγκο και την ταχύτητα του αέρα που περνάει από τους πνεύμονες στη γλώσσα. Έχει δειχθεί ότι σε περιόδους συναισθηματικής φόρτισης ασκείται μεγαλύτερη πίεση απ'οτι στον ουδέτερο λόγο κατά το άνοιγμα και το κλείσιμο των φωνητικών χορδών [52].

Μέσω του παλμού των φωνητικών χορδών, υπολογίζεται η διάρκεια του ανοίγματος και κλεισίματος των φωνητικών χορδών, ο λόγος και η διαφορά τους, καθώς και η πρώτη παράγωγός τους. Παρατηρείται ότι η χαρά έχει μεγαλύτερης διάρκειας περίοδο ανοίγματος

σε σχέση με τη λύπη και το θυμό [37]. Αντίθετα στο συναίσθημα της κατάθλιψης, γίνεται μεγαλύτερη προσπάθεια να περάσει ο αέρας από τις φωνητικές χορδές, οπότε υπάρχει μικρότερη περίοδος ανοίγματος και μεγαλύτερη περίοδος κλεισίματος των φωνητικών χορδών [52]. Επίσης, το φάσμα του παλμού των φωνητικών χορδών μπορεί να διαχωρίσει τον αγχώδη από τον ήρεμο λόγο [35]. Απότομες μεταβολές και οξείες γωνίες του φάσματος αυτού αντιστοιχούν σε αγχώδη ομιλία, ενώ ομαλές μεταβολές σε ήρεμο και ουδέτερο λόγο.

Περιοχές Φωνητικού Σωλήνα

Κατά την έκφραση συναισθήματος μεταβάλλεται το σχήμα του φωνητικού σωλήνα [36, 95], το οποίο σχετίζεται με την άρθρωση του λόγου. Έχει παρατηρηθεί ότι η άρθρωση είναι τεταμένη στο θυμό, μπερδεμένη στη θλίψη, ακριβής στο φόβο και παρόμοια με τον κανονικό λόγο στο συναίσθημα της χαράς και της αποστροφής [39, 15].

Ο φωνητικός σωλήνας προσομοιάζεται με σύνθεση ορθογωνίων σωλήνων συγκεκριμένου μήκους και μεταβλητού εμβαδού πλευράς [36, 95]. Για τη διαφοροποίηση του άγχους από το ουδέτερο συναίσθημα υπολογίζονται τα πηλίκα των εμβαδών των διαφόρων σωλήνων και βρίσκεται η περιβάλλουσά τους στο χρόνο. Παρατηρείται ότι στην ουδέτερη ομιλία ο λόγος του εμβαδού των σωλήνων που τοποθετούνται στο πίσω μέρος της φαρυγγικής κοιλότητας προς το εμβαδόν αυτών στο μπροστινό μέρος του στόματος είναι μεγάλος, που υπονοεί αύξηση του όγκου της φαρυγγικής κοιλότητας και μείωση του όγκου στο στόμα. Ακριβώς το αντίθετο συμβαίνει με την αγχώδη φωνή. Έχει βρεθεί ότι τα χαρακτηριστικά του φωνητικού σωλήνα διαχωρίζουν με μεγάλη επιτυχία το άγχος από το ουδέτερο συναίσθημα.

Χαρακτηριστικά LPC(Linear Predictive coding)

Τα LPC χαρακτηριστικά είναι από τις πιο βασικές παραμέτρους κωδικοποίησης της φωνής και εμπεριέχουν πληροφορία για το pitch, τα formants, το φάσμα φωνής και το φωνητικό σωλήνα. Στην αναγνώριση συναισθήματος έχουν χρησιμοποιηθεί οι 14 πρώτοι LPC συντελεστές σε διανύσματα που περιείχαν ενέργεια, διάρκεια, pitch και jitter [66].

Χαρακτηριστικά LPCC(Linear Predictive coding cepstrum)

Τα χαρακτηριστικά LPCC αντικατοπτρίζουν το σχήμα του φωνητικού σωλήνα και έχουν χρησιμοποιηθεί περιορισμένα στην αναγνώριση συναισθήματος [82, 97]. Σε σύγκριση με τα χαρακτηριστικά ενέργειας LFPC έχει παρατηρηθεί ότι έχουν χειρότερη αποδοτικότητα [57].

Φασματική Ισορροπία

Το χαρακτηριστικό αυτό [96] απεικονίζει ένα σταθμισμένο μέσο όρο της στιγμιαίας συχνότητας του σήματος με βάρη το φασματικό περιεχόμενο του πλάτους και υπολογίζεται στα έμφωνα τμήματα του λόγου. Παρατηρείται ότι η μέση τιμή της φασματικής ισοροπίας είναι αυξημένη για το θυμο, τη χαρά και τη λύπη σε σχέση με το ουδέτερο. Αυτό μπορεί να δικαιολογηθεί από το γεγονός ότι κατά την έκφραση των συναισθημάτων αυτών περνάει πιο πολύς αέρας από τις φωνητικές χορδές οπότε η φωνή έχει μεγαλύτερη ενέργεια στις υψηλές συχνότητες.

Χαρακτηριστικά Τονικότητας

Τα χαρακτηριστικά αυτά αντιπροσωπεύουν την τονικότητα στα όρια των φράσεων καθώς και στις συλλαβές [3]. Πιο συγκεκριμένα, μετράται ο τονισμός του pitch σε περιοχές υψηλού και χαμηλού pitch, καθώς και ο τονισμός στα μέσα της φράσης. Αποδείχθηκε ότι

σε σύγκριση με το κλασικό pitch, τα χαρακτηριστικά αυτά μπορεί να δώσουν καλύτερα αποτελέσματα αναγνώρισης και ο συνδυασμός των ToBI με το pitch μπορεί να είναι ακόμα πιο αποδοτικός [38].

Cepstral Χαρακτηριστικά

Σε πολλές έρευνες για αναγνώριση συναισθήματος έχουν χρησιμοποιηθεί cepstral χαρακτηριστικά, που είναι ευρέως διαδεδομένα στην αυτόματη αναγνώριση λόγου, γιατί φαίνεται να υπερτερούν των γραμμικών χαρακτηριστικών. Τα cepstral χαρακτηριστικά ανταναklούν τη φασματική δομή της φωνης στην κλίμακα Mel, που είναι η κλίμακα της ανθρώπινης ακουστικής [90].

Για την κατηγοριοποίηση συναισθημάτων θυμού, αποστροφής, φόβου, χαράς, λύπης, έκπληξης και ουδέτερου στην ασαμική γλώσσα χρησιμοποιήθηκαν MFCC (Mel frequency Cepstral Coefficients) χαρακτηριστικά με GMM ταξινομητή και ποσοστό επιτυχίας 73.4%. Για το ίδιο πρόβλημα εισάχθηκαν και MFCC χαρακτηριστικά των οποίων η ενέργεια (πριν λογαριθμηθεί) υπολογίζεται με βάση τον Teager τελεστή. Αυτά ονομάζονται tfMFCC χαρακτηριστικά και επιτυγχάνουν 75.1% αναγνώριση [43].

Σε άλλη έρευνα [93] με σκοπό την αναγνώριση χαράς, λύπης, θυμού, φόβου, έκπληξης και αποστροφής, διαπιστώθηκε ότι τα MFCC χαρακτηριστικά αν προστεθούν στα χαρακτηριστικά της προσωδίας κάνουν πιο αποτελεσματική την αναγνωριση συναισθήματος μέσω K-nearest Neighbors και Fisher's Linear Discriminant Analysis, ενώ δυσχεραίνουν την αναγώριση μέσω GMM, Maximum Likelihood και νευρωνικού δικτύου.

Τα MFCC χαρακτηριστικά έχουν συνδυαστεί και με τη μέθοδο Maximum Likelihood, με την οποία επιχειρήθηκε η ταξινόμηση λύπης, χαράς, θυμού και ουδέτερου. Διαπιστώθηκε [9] ότι πιο αποτελεσματικά από τα MFCC χαρακτηριστικά είναι τα MFB, που αποτελούν τη λογαριθμημένη ενέργεια της φωνής σε φίλτρα τοποθετημένα σύμφωνα με την κλίμακα Mel. Ο λόγος είναι ότι τα MFB χαρακτηριστικά δεν αλλοιώνουν τις ακουστικές διαφορές μεταξύ του λόγου που περιέχει συναίσθημα και του ουδέτερου λόγου, πράγμα που γίνεται στα MFCC μέσω του διακριτού μετασχηματισμού συνημιτόνου. Επίσης, σε επόμενη έρευνα [44] παρατηρήθηκε ότι τα MFB χαρακτηριστικά δίνουν ίδια αποτελέσματα ταξινόμησης με μικρότερη διαστασιμότητα απ'οτι τα MFCC χαρακτηριστικά.

Για την αναγνώριση των διαφόρων ειδών αγχώδους λόγου έχουν χρησιμοποιηθεί χαρακτηριστικά MFCC, δέλτα MFCC, δέλτα-δέλτα MFCC, αυτοσυσχέτιση των MFCC και ετεροσυσχέτιση των MFCC [36, 95]. Θεωρείται ότι τα χαρακτηριστικά αυτά δίνουν πληροφορία σχετικά με τις αλλαγές στη φασματική δομή της φωνής και στο φωνητικό σωλήνα που προέρχονται από το άγχος. Η αυτοσυσχέτιση των MFCC μοντελοποιεί τη σχετική ενέργεια μεταξύ των καναλιών και των παραθύρων του σήματος φωνής και υποστηρίζεται ότι είναι το πιο αποτελεσματικό χαρακτηριστικό για την κατηγοριοποίηση του αγχώδους λόγου. Σε μετέπειτα έρευνα [68] βρέθηκε ότι η ενίσχυση των ESA χαρακτηριστικών από MFCC επιφέρει βελτίωση μέχρι και 12.99% για την κατηγοριοποίηση άγχους.

Χαρακτηριστικά Διαμόρφωσης AM-FM

Τα χαρακτηριστικά αυτά είναι μη γραμμικά, αφορούν τις διαμορφώσεις πλάτους - συχνότητας των συντονισμών της φωνής ή και του pitch, έχουν χρησιμοποιηθεί κυρίως στην αναγνώρισδη άγχους και προέρχονται από τον Energy Separation Algorithm(ESA) [51]. Ο ESA είναι ένας ταχύς και αποδοτικός αλγόριθμος για την ανίχνευση των χαρακτηριστικών αυτών

με βασικό τελεστή τον Teager-Energy-Operator (TEO) [41]. Αυτός μπορεί να εντοπίσει τη διαμόρφωση του σήματος φωνής και να χωρίσει το σήμα σε AM και FM μέρη. Το FM κομμάτι του σήματος φωνής αντανακλά τις απότομες μεταβολές των formants και του pitch, που είναι πιο εμφανείς στον αγχώδη απ'οτι στον ουδέτερο λόγο. Το AM κομμάτι δείχνει τις απότομες μεταβολές του πλάτους των συντονισμών της φωνής και του πλάτους της διέγερσης.

TEO-Auto-Env, TEO-Pitch και TEO-CB-Auto-Env

Σε έρευνες [98, 99] εισάγονται δύο νέα χαρακτηριστικά που βασίζονται στον TEO τελεστή. Αρχικά, το TEO-Auto-Env χαρακτηριστικό υπολογίζεται από το εμβαδό της περιβάλλουσας της αυτοσυσχέτισης του AM κομματιού. Το πλεονέκτημα του χαρακτηριστικού αυτού είναι ότι αποσιωπά τις πολύ γρήγορες μεταβολές και παράλληλα διατηρεί τις μεταβολές στο λόγο που οφείλονται στο άγχος. Το TEO-Pitch χαρακτηριστικό βασίζεται σε μία συνάρτηση αυτοσυσχέτισης και σε δυναμικό προγραμματισμό και αντανακλά η δομή του pitch στη φωνή. Βρέθηκε ότι το TEO-Pitch επιτυγχάνει καλύτερα αποτελέσματα για την κατηγοριοποίηση 4 ειδών αγχώδους φωνής σε σχέση με το pitch, το TEO-Auto-Env χαρακτηριστικό και το FM κομμάτι της φωνής που προκύπτει από τον TEO τελεστή. Τέλος, το TEO-CB-Auto-Env χαρακτηριστικό βρίσκεται με τον ίδιο τρόπο με το TEO-Auto-Env με τη μόνη διαφορά ότι υπολογίζεται σε 16 ζώνες συχνοτήτων από 0 έως 3700 Hz όχι ομοιόμορφα κατανεμημένες που θεωρούνται οι πιο αντιπροσωπευτικές της ανθρώπινης ακοής. Το χαρακτηριστικό αυτό φαίνεται να είναι πιο αποδοτικό για την κατηγοριοποίηση ουδέτερης, θυμωμένης, δυνατής και Lombard ομιλίας σε σχέση με το MFCC και το pitch.

3.2.2 Αλγόριθμοι Επιλογής Χαρακτηριστικών

Πριν από την ταξινόμηση των χαρακτηριστικών που υπολογίζονται μπορεί να προηγηθεί ένα επιπλέον βήμα, αυτό της επιλογής των βέλτιστων χαρακτηριστικών από ένα γενικό σύνολο. Οι αλγόριθμοι επιλογής χαρακτηριστικών εκτελούν μία αναζήτηση ώστε να βρουν ένα βέλτιστο υποσύνολο (ή κοντά στο βέλτιστο) χαρακτηριστικών με βάση ένα συγκεκριμένο κριτήριο. Η επιλογή χαρακτηριστικών είναι ευεργετική για την ταξινόμηση για τους επόμενους λόγους [76]:

- Χαρακτηριστικά που είναι τελειώς ασυσχέτιστα μεταξύ τους ή που δε δίνουν καμία πληροφορία μπορεί να αποτελέσουν πηγή αστάθειας για έναν ταξινομητή.
- Μεγάλος αριθμός χαρακτηριστικών απαιτεί και μεγάλο αριθμό παρατηρήσεων, που δεν είναι πάντα διαθέσιμες.
- Εξαλείφοντας χαρακτηριστικά που δεν προσδίδουν καμία πληροφορία, μειώνεται ο χρόνος ταξινόμησης και συλλογής των χαρακτηριστικών.

Στη συνέχεια παρουσιάζουμε τους πιο κοινούς αλγόριθμους επιλογής χαρακτηριστικών που έχουν χρησιμοποιηθεί στην αναγνώριση συναισθήματος.

Linear Discriminant Analysis

Στόχος της μεθόδου LDA είναι η επιλογή των χαρακτηριστικών που αυξάνουν τη διαταξική διασπορά μεταξύ των κλάσεων των συναισθημάτων, ενώ παράλληλα μειώνουν την ενδοταξική διασπορά σε κάθε κλάση. Στην έρευνα [75] χρησιμοποιήθηκε η LDA για την επιλογή ενός

διανύσματος διάστασης 33 μέσα από 200 συνολικά χαρακτηριστικά. Αποδεικνύεται μάλιστα ότι τα χαρακτηριστικά του pitch κατείχαν τις πρώτες θέσεις στην επιλογή χαρακτηριστικών.

RELIEF-F

Ο αλγόριθμος RELIEF-F παίρνει τυχαία δείγματα από τα δεδομένα, βρίσκει του κοντινότερους γείτονές τους και ρυθμίζει το βάρος κάθε χαρακτηριστικού έτσι ώστε να δίνει πιο πολύ βάρος σε χαρακτηριστικά που ξεχωρίζουν το τυχαίο δείγμα από γείτονές του διαφορετικών κλάσεων. Έχει χρησιμοποιηθεί σε έρευνα για αναγνώριση συναισθήματος [61] ξεχωρίζοντας το μέγιστο, τυπική απόκλιση, εύρος και μέσο όρο του pitch ως τα πιο διακριτά χαρακτηριστικά.

Σειριακές Τεχνικές Αναζήτησης (Sequential Search Strategies)

Οι σειριακές τεχνικές αναζήτησης προσθέτουν ή αφαιρούν σειριακά χαρακτηριστικά στο υποσύνολο. Το μειονέκτημα των αλγορίθμων αυτών είναι ότι μπορεί να παγιδευτούν σε τοπικά ελάχιστα. Οι μέθοδοι σειριακής αναζήτησης που θα μελετήσουμε στην αναγνώριση συναισθήματος είναι: Σειριακή αναζήτηση προς τα εμπρός (Sequential Forward Selection), Σειριακή αναζήτηση 2 κατευθύνσεων, Κυμαινόμενη Σειριακή Αναζήτηση προς τα εμπρός (Sequential Floating Forward Selection).

Ο SFS είναι ο πιο απλός αλλά άπληστος (greedy) αλγόριθμος σειριακής αναζήτησης, καθώς αρχίζοντας από ένα κενό σύνολο διαλέγει κάθε φορά το χαρακτηριστικό που μεγιστοποιεί το κριτήριο αναζήτησης. Έχει χρησιμοποιηθεί στις έρευνες [18] και [77].

Στη σειριακή αναζήτηση 2 κατευθύνσεων, γίνονται 2 αναζητήσεις. Η πρώτη (προς τα εμπρός αναζήτηση) ξεκινάει από ένα κενό σύνολο προσθέτοντας κάθε φορά το χαρακτηριστικό που βελτιστοποιεί το κριτήριο αναζήτησης, ενώ η δεύτερη (προς τα πίσω αναζήτηση) ξεκινάει από το αρχικό σύνολο χαρακτηριστικών και αφαιρεί κάθε φορά το χαρακτηριστικό που έχει τη χειρότερη επίδραση. Στην έρευνα [46] η προς τα εμπρός και προς τα πίσω αναζητήσεις έδωσαν παρόμοια αποτελέσματα και έδειξαν ότι το pitch και η ενέργεια είναι τα πιο σημαντικά χαρακτηριστικά στην αναγνώριση του αγχώδους λόγου από τον ουδέτερο. Παρόμοια αναζήτηση χρησιμοποιήθηκε και στο [93], αλλά με κριτήριο την απόσταση Mahalanobis μεταξύ των κλάσεων.

Η μέθοδος SFFS είναι η πιο περίπλοκη μέθοδος αναζήτησης. Σύμφωνα με αυτή, ξεκινώντας από το κενό σύνολο επιλέγεται ένα χαρακτηριστικό και στη συνέχεια αφαιρούνται χαρακτηριστικά μέχρι το κέρδος της συνάρτησης κριτηρίου να αρχίζει να μειώνεται, οπότε και ξαναγίνεται προσθήκη χαρακτηριστικού. Η διάσταση του συνόλου που δημιουργείται είναι 'κυμαινόμενη', για αυτό και το όνομα της μεθόδου. Έχει χρησιμοποιηθεί σε έρευνες στην αναγνώριση συναισθήματος με συνάρτηση κριτηρίου τα ποσοστά επιτυχίας δικτύου SVM [72] και μπεϋζιανού ταξινομητή με γκαουσιανή κατανομή της υπό συνθήκη πιθανότητας [88]. Η SFFS μπορεί να εφαρμοστεί και σε συνδυασμό με LDA [62]. Πρώτα διαλέγονται τα πιο αποδοτικά χαρακτηριστικά μέσω της SFFS και στη συνέχεια γίνεται μείωση της διαστασιμότητας μέσω της LDA. Τέλος, προτάθηκε και μία βελτιωμένη έκδοση του αλγορίθμου SFFS που δίνει παρόμοια αποτελέσματα με μειωμένο υπολογιστικό κόστος, αποφεύγοντας τις περιττές επαναλήψεις [91].

Γενετικοί Αλγόριθμοι

Οι γενετικοί αλγόριθμοι είναι αλγόριθμοι τυχαίας αναζήτησης που εστιάζουν στην επιλογή, μετάλλαξη και ανασυνδυασμό χαρακτηριστικών. Βασίζονται στην αρχή του Δαρβίνου για την επιβίωση των επικρατέστερων χαρακτηριστικών.

Σε μία έρευνα [76] προτάθηκε η χρήση γενετικού αλγορίθμου για την εύρεση των χειρότερων χαρακτηριστικών και προτάσεων με βάση ένα κριτήριο ποικιλίας του συνολικού πληθυσμού. Σε κάθε επανάλυψη του αλγορίθμου ο πληθυσμός αποτελείται από ορισμένο πλήθος χρωμοσωμάτων N_p, στα οποία γίνονται ανταλλαγές και μεταλλάξεις με πιθανότητα που εξαρτάται από την ποικιλία του. Στη συνέχεια προστίθενται στον πληθυμό N_p επιπλέον χρωμοσώματα και επαναυπολογίζεται το κριτήριο ποικιλίας για τον πληθυσμό με τα $2N_p$ χρωμοσώματα. Αφαιρούνται τα N_p χρωμοσώματα που επιφέρουν τη μέγιστη ποικιλία (δηλαδή τα $N - p$ καλύτερα χρωμοσώματα) και επαναλαμβάνεται η παραπάνω διαδικασία, μέχρι να μελετηθούν όλα τα χρωμοσώματα ή να μηδενιστεί η ποικιλία του πληθυσμού. Στον παραπάνω αλγόριθμο ως χρωμοσώματα μπορεί να θεωρηθούν τα χαρακτηριστικά αλλά και οι προτάσεις των διαφόρων συναισθημάτων. Με τον τρόπο αυτό, από χαρακτηριστικά pitch, διαμορφωτριών συχνοτήτων, ενέργειας και φάσματος, επιλέγεται το χειρότερο, που είναι η μέση τιμή της 2ης διαμορφώτριας συχνότητας. Επίσης, από όλες τις προτάσεις επιλέγεται η χειρότερη. Το αποτέλεσμα είναι η βελτίωση της αναγνώρισης με μπεϋζιανό ταξινομητή από 47.06% σε 48.91% για 4 συναισθήματα (θυμός, χαρά, λύπη, έκπληξη) και το ουδέτερο.

Η χρήση γενετικού αλγορίθμου έχει γίνει και για την επιλογή όχι μόνο των χειρότερων αλλά και των καλύτερων χαρακτηριστικών [73]. Από ένα σύνολο χαρακτηριστικών που αφορούν την ένταση, τον τονισμό, τη διάρκεια , την επιμήκυνση, τις διαμορφώτριες συχνότητες (formants), και το φάσμα του σήματος, έχει γίνει μία προεπιλογή μέσω του SFFS (Sequential Forward Floating Search) αλγορίθμου και στην συνέχεια ένα 'ραφινάρισμα' μέσω γενετικού αλγορίθμου. Ο γενετικός αλγόριθμος εφαρμόζει επιλογή χρωμοσωμάτων μέσω του μηχανισμού της ρουλέτας, όπου κάθε χρωμόσωμα παραμένει στον πληθυσμό με πιθανότητα ανάλογη της ποικιλίας του πληθυσμού. Η διασταύρωση γίνεται τέμνοντας το χρωμόσωμα κάθε γονέα περίπου στη μέση και ενώνοντας τα 2 τμήματα με τα αντίστοιχα τμήματα άλλου γονέα. Τέλος, κατά τη μετάλλαξη αλλάζει η κατάσταση ενός γονιδίου (που υποδηλώνει την ύπαρξη ή μη χαρακτηριστικού) με πιθανότητα 0.5. Έπειτα υπολογίζεται η συνάρτηση ποικιλίας από το ποσοστό επιτυχίας ενός SVM ταξινομητή και επαναλαμβάνεται η παραπάνω διαδικασία για ορισμένο αριθμό φορών. Παρατηρήθηκε ότι ο γενετικός αλγόριθμος βελτίωσε τα ποσοστά αναγνώρισης σε τρεις βάσεις συναισθηματικής ομιλίας σε σχέση με την εφαρμογή μόνο του SFFS, όπως φαίνεται στον παρακάτω πίνακα.

μέθοδος	DES	EMO-DB	EA-CAR
SFFS	74.15	87.50	75.08
Γενετικός	76.15	88.82	77.18

3.2.3 Αλγόριθμοι Μείωσης Διαστασιμότητας

Η μείωση διαστασιμότητας είναι ένα κοινό πρόβλημα στην ψηφιακή επεξεργασία σήματος και προέρχεται από την ανάγκη για περιορισμό του όγκου των δεδομένων. Ιδιαίτερα τα χαρακτηριστικά που χρησιμοποιούνται στην αναγνώριση συναισθήματος σχετίζονται μεταξύ τους και πρέπει να βρεθούν τρόποι εξάλειψης της περιττής πληροφορίας.

Η μέθοδος PCA είναι η πιο συχνά χρησιμοποιούμενη τεχνική για το σκοπό αυτό και έχει εφαρμοστεί και στην αναγνώριση συναισθήματος [86].

Μία πιο σύνθετη μέθοδος είναι ο ισομετρικός μετασχηματισμός, που αποτελεί έναν από τους πιο δημοφιλείς τρόπους εύρεσης της πολλαπλής πληροφορίας (manifold). Σύμφωνα με τη μέθοδο αυτή [80], βρίσκονται τα γειτονικά δείγματα i, j, δηλαδή τα σημεία που δεν ξεπερνούν μία συγκεκριμένη απόσταση $d(i,j) < \epsilon$. Οι γειτονικές αυτές σχέσεις αναπαρίστανται με ένα γράφο G με κορυφές τα σημεία δειγμάτων και ακμές:

$$d = \left\{ \begin{array}{c} d(i,j), \text{ αν τα i,j είναι γειτονικά} \\ \infty, \text{ αλλιώς} \end{array} \right\}$$

Στη συνέχεια, κατασκευάζεται ο γράφος G_M με ακμές τις ελάχιστες αποστάσεις μεταξύ δύο σημείων, έστω $D_G = \{d_M(i,j)\}$ οι αποστάσεις του γράφου. Τέλος βρίσκεται ένας ευκλείδιος χώρος Y διάστασης d που διατηρεί τη δομή του αρχικού χώρου και του οποίου τα διανύσματα $\mathbf{y}_i$ ελαχιστοποιούν τη συνάρτηση κόστους:

$$E = ||\tau(D_G) - \tau(D_Y)||_{L^2}$$

όπου D_Y είναι ο πίνακας των ευκλείδιων αποστάσεων των σημείων το $||A||_{L^2}$ αναπαριστά την L^2 απόσταση $\sqrt{\Sigma_{ij} A_{ij}}$. Ο τελεστής τ μετατρέπει τις αποστάσεις σε εσωτερικά γινόμενα, δηλαδή $\tau(D) = -HSH/2$ όπου S είναι η απόσταση των τετραγωνικών αποστάσεων $S_{ij} = D_{ij}^2$ και $H = D_{ij} - 1/N$.
Η τεχνική αυτή εφαρμόζεται σε χαρακτηριστικά MFCC, MFB και χαρακτηριστικά άρθρωσης [44]. Η ταξινόμηση γίνεται με GMMs και το αποτέλεσμα που προκύπτει από τον ισομετρικό μετασχηματισμό (σε χώρο 3-5 διαστάσεων) επιτυγχάνει παρόμοια ποσοστά αναγνώρισης με το αποτέλεσμα χωρίς το μετασχηματισμό αυτό. Επίσης, η μείωση αυτή της διαστασιμότητας κρατάει μεγαλύτερη πληροφορία στα χαρακτηριστικά άρθρωσης και MFB απ΄ ότι στα MFCC.

3.2.4 Τεχνικές Κατηγοριοποίησης

Με βάση τα παραπάνω χαρακτηριστικά, επιχειρείται η αναγνώριση των συναισθημάτων και των ειδών του αγχώδους λόγου. Χωρίζουμε τις τεχνικές κατηγοριοποίησης σε αυτές που εκμεταλλεύονται τη χρονική ακολουθία των χαρακτηριστικών και σε αυτές στις οποίες η χρονική πληροφορία δε λαμβάνεται υπόψη.

K-μέσοι (K-means), K-Κοντινότεροι Γείτονες (K-Nearest Neighbours)

Η κατηγοριοποίηση με k-μέσους και k κοντινότερους γείτονες είναι από τις πιο απλές μεθόδους και έχουν χρησιμοποιηθεί πολύ στην αναγνώριση συναισθήματος [18, 31, 61, 93].

Με τη μεθοδο των k κοντινότερων γειτόνων επιχειρήθηκε η ταξινόμηση 6 συναισθημάτων (χαρά, λύπη, θυμός, φόβος, έκπληξη και αποστροφή) με διάνυσμα διάστασης 55 που περιέχει χαρακτηριστικά προσωδίας, MFCC και διαμορφωτριών συχνοτήτων (formants). Η μέθοδος αυτή έδωσε αποτέλεσμα 62.28%, που ξεπέρασε στο συγκεκριμένο πείραμα το ποσοστό αναγνώρισης των GMMs 1-3 γκαουσιανών για κάθε συναίσθημα καθώς και του νευρωνικού δικτύου. Ελαφρώς βελτιωμένο αποτέλεσμα επιτεύχθηκε μέσω της γραμμικής διακριτής ανάλυσης του Fisher με ποσοστό 64.44% [93].

Η αποτελεσματικότητα της ταξινόμησης με τους k-κοντινότερους γείτονες αποδεικνύεται, επίσης, σε σχέση με την ταξινόμηση μεγίστης πιθανοφάνειας με μπευζιανό κατηγοριοποιητή και την ταξινόμηση με γκαουσιανό πυρήνα. Επιτυγχάνεται λάθος ταξινόμησης 32% με χαρακτηριστικά που αφορούν το ομαλοποιημένο pitch και την παράγωγό του, το ρυθμό ομιλίας και τα έμφωνα τμήματα φωνής [18].

Κατηγοριοποίηση με k-κοντινότερους γείτονες έχει χρησιμοποιηθεί και στην αναγνώριση του άγχους, όπου επιχειρείται η δυαδική αναγνώριση του ουδέτερου λόγου από το θυμωμένο, το δυνατό και το Lombard. Τα μικρότερα λάθη προκύπτουν με βάση διάνυσμα διάστασης 10 που περιείχε χαρακτηριστικά pitch, διάρκειας και έντασης και είναι (1.67%, 0%), (3.03%, 3.03%) και (13.64%, 16.67%) για τα παραπάνω ζευγάρια ειδών ομιλίας αντίστοιχα [35].

Σε μία άλλη έρευνα [31] επιχειρήθηκε η κωδικοποίηση των διανυσμάτων εισόδου μέσω του αλγορίθμου των k-μέσων. Δημιουργήθηκαν δύο κωδικοποιήσεις ανάλογα με το φύλο του εκφωνητή από τον οποίο προέρχονταν τα εκάστοτε δεδομένα. Για την αναγνώριση χρησιμοποιήθηκαν τα MFCC χαρακτηριστικά υπολογισμένα σε 4 ζώνες συχνοτήτων που καθορίστηκαν από το μετασχηματισμό κυματιδίων (wavelets). Για τον έλεγχο του συστήματος δεδομένης μίας φράσης εισόδου, υπολογιζονται τα χαρακτηριστικά εισόδου σε κάθε ζώνη συχνοτήτων, συγκρίνονται με τις τιμές κωδικοποίησης και βρίσκονται τα ποσοστά ταιριάσματος σε κάθε κλάση. Τα ποσοστά ταιριάσματος κάθε κλάσης αθροίζονται ώστε να προκύψει το τελικό ποσοστό ταιριάσματος. Το συναίσθημα που αναγνωρίζεται είναι αυτό για το οποίο προκύπτει το μέγιστο συνολικό ποσοστό ταιριάσματος:

$$e_i = arg \max_{i=1,\ldots,N} \sum_{k=1}^{4} MatchingScore_k$$

όπου τα e_i για $i = 1, \ldots, N$ είναι τα συνασθήματα υπό αναγνώριση και το $MatchingScore_k$ για $k = 1, \ldots, 4$ είναι το ποσοστό ταιριάσματος του διανύσματος εισόδου με τις κωδικοποιημένες τιμές για κάθε ζώνη συχνοτήτων. Χρησιμοποιώντας τις τρεις χαμηλόρες ζώνες συχνοτήτων για τα 6 βασικά συναισθήματα (χαρά, λύπη, θυμός, έκπληξη, φόβος και απαρέσκεια) επιτυγχάνεται ποσοστό αναγνώρισης 93.3% για τους άντρες και 86% για τις γυναίκες, ενώ σε όλες τις ζώνες συχνοτήτων τα ποσοστά επιτυχίας γίνονται 85% και 93.3% αντίστοιχα. Τα διαφορετικά ποσοστά σε διαφορετικές ζώνες συχνοτήτων δείχνουν ότι το συναίσθημα εκφράζεται με διαφορετικό τρόπο στους άντρες και τις γυναίκες.

Μείγματα Γκαουσιανών (GMMs)

Στο γκαουσιανό μαρκοβιανό μοντέλο θεωρείται ότι τα δεδομένα μιας κλάσης μπορούν να αναπαρασταθούν από μία ή περισσότερες γκαουσιανές κατανομές. Τα GMMs στην αναγνώριση συναισθήματος έχουν χρησιμοποιηθεί με βάση ποικίλα χαρακτηριστικά [40, 42, 68].

Για την κατηγοριοποίηση 6 συναισθημάτων (θυμός, αποστροφή, φόβος, χαρά, λύπη και έκπληξη) με GMM 12 καταστάσεων και με βάση MFCC χαρακτηριστικά επιτυγχάνεται ποσοστό 74.4% [42]. Χρησιμοποιώντας το TEO-Auto-Env για GMM 32 καταστάσεων με σκοπό την αναγνώριση του άγχους στη φωνή, το ποσοστό επιτυχίας γίνεται 63.71% [68]. Στο ίδιο πείραμα μάλιστα αποδεικνύεται ότι παρόμοια ποσοστά αναγνώρισης έχουμε αν διπλασιάσουμε τον αριθμό των γκαουσιανών στο GMM. Επίσης, για την εκπαίδευση του μοντέλου αρκούν 12 ομιλητές, καθώς για περισσότερους από 12 μέχρι και 35 ομιλητές τα ποσοστά επιτυχίας παραμένουν ίδια. Τέλος, η εκπαίδευση των GMMs με διάνυσμα διάστασης 21 που περιέχει χαρακτηριστικά pitch, ενέργειας και διάρκειας δίνει αποτέλεσμα 68.6%.

Τα GMMs έχουν χρησιμοποιηθεί και για την ταξινόμηση των συναισθημάτων χαράς, λύπης και θυμού με βάση χαρακτηριστικά φωνητικών χορδών [37]. Όπως είπαμε και προηγουμένως, τα χαρακτηριστικά αυτά περιλαμβάνουν το άνοιγμα και κλείσιμο των φωνητικών χορδών, το λόγο των δύο μεγεθών, τη διαφορά τους καθώς και την 1η παράγωγό τους. Η κατηγοριοποίηση των τριών συναισθημάτων με βάση το σύστημα αυτό έγινε με ποσοστό επιτυχίας από 48.96% έως και 54.17%. Οι ίδιοι ερευνητές μελέτησαν το συνδυασμό χαρακτηριστικών pitch με χαρακτηριστικά τονικότητας. Για την ταξινόμηση των τριών παραπάνω συναισθημάτων με GMMs βρέθηκαν ποσοστά επιτυχίας από 93.44% μέχρι και 100%.

Τέλος, έχει αποδειχθεί ότι το φύλο παίζει ρόλο στην κατηγοριοποίηση των συναισθημάτων με GMMs [88]. Για την ταξινόμηση 4 συναισθημάτων (θυμός, χαρά, λύπη, έκπληξη) και του ουδέτερου χρησιμοποιήθηκε ο αλγόριθμος SFFS (Sequential Floating Forward Selection) σε 87 συνολικά χαρακτηριστικά που αφορούσαν ενέργεια, ένταση και pitch. Το κριτήριο για την επιλογή των χαρακτηριστικών είναι η πιθανότητα σωστής κατηγοριοποίησης από μπευζιανό ταξινομητή που χρησιμοποιεί μείγμα γκαουσιανών. Βρίσκεται ότι για την κατηγοριοποίηση συναισθημάτων ξεχωριστά για τα δύο φύλα αρκεί μία μόνο γκαουσιανή για κάθε GMM με αποτέλεσμα 61.8% για τους άντρες και 57.6% για τις γυναίκες. Αντίθετα, αν θέλουμε να ταξινομήσουμε τα συναισθήματα και για τα δύο φύλα, πρέπει να χρησιμοποιήσουμε GMM 2 καταστάσεων, πράγμα που δείχνει ότι κάθε φύλο εισάγει και μία γκαουσιανή.

Κρυφά Μαρκοβιανά Μοντέλα (HMMs)

Τα κρυφά μαρκοβιανά μοντέλα χρησιμοποιούνται ευρέως στην αναγνώριση ομιλίας, οπότε έχουν υιοθετηθεί και στην αναγνώριση συναισθήματος λόγω της αποτελεσματικότητάς τους.

Στις περισσότερες έρευνες ένα HMM μοντελοποιεί μία κατηγορία συναισθήματος [57, 74]. Οι Nwe, Foo, De Silva μοντελοποιούν μία πρόταση από ένα εργοδικό HMM, το οποίο πλεονεκτεί σε σχέση με το left-right HMM. Αυτό γιατί το συναίσθημα έχει ακανόνιστη χρειά και δεν μπορεί να προσομοιωθεί με ακολουθιακά γεγονότα. Για παράδειγμα, αν μία παύση σχετίζεται με το συναίσθημα της λύπης, αυτή μπορεί να συμβεί σε οποιαδήποτε στιγμή μέσα στην πρόταση. Χρησιμοποιώντας εργοδικά HMMs 4 καταστάσεων το καθένα, με βάση LFPC χαρακτηριστικά, αναγνωρίζονται σωστά 6 συναισθήματα (θυμός, απαρέσκεια, φόβος, χαρά, λύπη και έκπληξη) με ποσοστό 78.1% ανεξαρτήτως πρότασης και ομιλητή.

Σε άλλη έρευνα [74] χρησιμοποιήθηκε 1 left-right HMM 64 καταστάσεων με μείγμα 4 γκαουσιανών ανά κατάσταση για κάθε συναίσθημα. Το διάνυσμα χαρακτηριστικών με τα οποίο έγινε η εκπαίδευση του HMM είναι 6 διαστάσεων και περιέχει το pitch, το λογάριθμο της μέσης ενέργειας για κάθε frame καθώς και την 1η και 2η παράγωγο των μεγεθών αυτών, όπως φαίνεται και στην παρακάτω σχέση.

$$m_i = (F_{0i}, \frac{dF_{0i}}{dt}, \frac{d^2F_{0i}}{dt^2}, E_i, \frac{dE_i}{dt}, \frac{d^2E_i}{dt^2})$$

όπου i είναι ο δείκτης των frames.
Ο μέσος όρος του ποσοστού επιτυχίας είναι 77.8% για 6 συναισθήματα (θυμός, απαρέσκεια, φόβος, έκπληξη, χαρά, λύπη) και το ουδέτερο.

Τέλος, σε παρόμοια έρευνα που έχει διεξαχθεί για την κατηγοριοποίηση των ειδών του άγχους έχει βρεθεί 89% ποσοστό επιτυχίας για την ταξινόμηση 3 ειδών άγχους και του ουδέτερου με βάση το TEO-Auto-Env χαρακτηριστικό.

Υπάρχουν ενδείξεις ότι τα συναισθήματα συμβάλουν στη διαμόρφωση των φωνημάτων στην ομιλία με διαφορετικούς τρόπους [48]. Έχει δειχθεί για παράδειγμα ότι η θέση της πρώτης διαμορφώτριας συχνότητας (formant) F_1 επηρεάζεται περισσότερο από το συναισθηματικό χρωματισμό για το φωνήεν /AA/, ενώ η θέση της δεύτερης διαμορφώτριας συχνότητας F_2 επηρεάζεται πιο πολύ για το φωνήεν /IY/. Έτσι, για την αναγνώριση συναισθήματος έχουν χρησιμοποιηθεί HMMs για κλάσεις φωνημάτων με βάση MFCC χαρακτηριστικά. Πιο συγκεκριμένα, χρησιμοποιήθηκαν 46 φωνήματα μίας συλλαβής ανεξάρτητα περιεχομένου από τη βάση δεδομένων TIMIT, τα οποία χωρίστηκαν σε 5 ομάδες: φωνήεντα (vowel), ένρινα (nasal), συρριστικά (stop), γλωττιδικά (glottal) και τριβόμενα (fricative). Κάθε κλάση φωνήματος αντιπροσωπεύεται από ένα HMM 3 καταστάσεων με 16 γκαουσιανές ανά κατάσταση. Για 3 συναισθήματα (θυμός, χαρά, λύπη) και το ουδέτερο, με βάση 13 MFCC χαρακτηριστικά καθώς και την 1η και 3η παράγωγό τους, επιτεύχθηκε 75.57% σωστή κατηγοριοποίηση. Ο συνδυασμός του παραπάνω συστήματος με κατηγοριοποιητή SVM για χαρακτηριστικά προσωδίας απέδωσε 76.12% επιτυχία.

Παρόμοια τεχνική έχει χρησιμοποιηθεί και για την κατηγοριοποίηση των ειδών του άγχους [98, 99]. Για κάθε είδος αγχώδους λόγου και για τα φωνήεντα κάθε λέξης εκπαιδεύται ένα HMM μοντέλο. Το ποσοστό αναγνώρισης για μη ανεξάρτητο περιεχόμενο λέξεων με βάση το TEO-Auto-Env χαρακτηριστικό είναι 92.9%, ενώ με βάση το MFCC είναι 90.9%.

Η ταξινόμηση με HMMs μπορεί να ενσωματωθεί σε έναν αλγόριθμό K-means, δημιουργώντας N HMM μοντέλα, ένα για κάθε κλάση [28]. Έστω $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ το σύνολο των χρονικών δεδομένων, $\lambda_n^{(l)}$ για $n = 1, \ldots, N$ οι παράμετροι του $n^{\text{οστού}}$ HMM και $\hat{n}_k^{(l)} = arg\max_n P(\mathbf{x}^k|\lambda_n^{(l)})$ η κλάση που μεγιστοποιεί την πιθανότητα του k χρονικού δείγματος στην l επανάληψη. Η αρχικοποίηση γίνεται με τυχαία ανάθεση των δειγμάτων στις κλάσεις και υπολογισμό της αρχικής πιθανοφάνειας $P^{(0)} = \sum_k logP(\mathbf{x}^k)|\lambda_{\hat{n}_k}{}^{(0)}$. Στη συνέχεια, μέσω του Baum-Welch αλγορίθμου υπολογίζονται οι παράμετροι των HMMs λ_n^{l+1} και μέσω του αλγορίθμου Viterbi η κλάση $\hat{n}_k^{l+1}$ στην οποία ανήκει κάθε δείγμα. Με βάση τη νέα αναδιάταξη των δειγμάτων, υπολογίζεται η νέα πιθανοφάνεια $P^{(\lambda+1)} = \sum_k logP(\mathbf{x}^k)|\lambda_{\hat{n}_k}^{(\lambda+1)}$ και επαναλαμβάνεται η διαδικασία μέχρι η πιθανοφάνεια να συγκλίνει. Η μέθοδος αυτή χρησιμοποιείται στην ταξινόμηση 4 ειδών αγχώδους λόγου με βάση χαρακτηριστικά Teager ενέργειας και υπερτερεί έναντι όλων των άλλων μεθόδων που χρησιμοποιήθηκαν (απλά HMMs, Single Autoregressive HMMs [28], Factorial HMMs [28], SVMs νευρωνικά δίκτυα) δίνοντας 80.42% ποσοστό επιτυχίας για αναγνώριση εξαρτώμενη από το χρήστη και 51.22% για αναγώριση ανεξάρτητη από το χρήστη.

Νευρωνικά Δίκτυα

Νευρωνικά δίκτυα έχουν χρησιμοποιηθεί στην ταξινόμηση συναισθήματος καθώς και στην κατηγοριοποίηση του αγχώδους λόγου.

Ένα νευρωνικό δίκτυο μπορεί να εκπαιδευτεί για όλες τις συναισθηματικές καταστάσεις έχοντας τόσες εξόδους όσες και το πλήθος των συναισθημάτων που μελετώνται. Για την ταξινόμηση 4 συναισθημάτων (χαρά, θυμός, λύπη, φόβος) και του ουδέτερου εκπαιδεύτηκε ένα νευρωνικό δίκτυο 2 στρωμάτων με τη μέθοδο της ανάστροφης διάδοσης (back propagation) με βάση χαρακτηριστικά pitch, ενέργειας, διαμορφωτριών συχνοτήτων και διάρκειας. Το ποσοστό επιτυχίας του δικτύου αυτού είναι 65% [61]. Σε παρόμοια έρευνα για την κατηγοριοποίηση των ειδών του άγχους κατά τη διάρκεια της οδήγησης, χρησιμοποιήθηκε επίσης νευρωνικό δίκτυο 2 στρωμάτων με 10 και 4 μονάδες αντίστοιχα στο 1ο και 2ο κρυμ-

μένο επίπεδο και ο επαναυπολογισμός των παραμέτρων του έγινε με τη μέθοδο ανάστροφης διάδοσης. Το μέσο αποτέλεσμα ήταν 50.57% [28].

Για την αναγνώριση συναισθήματος έχει χρησιμοποιηθεί και ένα σύνολο από νευρωνικά δίκτυα [61]. Αυτό αποτελείτα από περιττό αριθμό δικτύων που έχουν εκπαιδευτεί σε διαφορετικά υποσύνολα δεδομένων. Η τελική απόφαση βασίζεται στο αποτέλεσμα της πλειοψηφίας των νευρωνικών δικτύων. Για 15 δίκτυα ο μέσος όρος του ποσοστού επιτυχίας είναι 70% για την κατηγοριοποίηση των 4 βασικών συναισθημάτων και του ουδέτερου.

Με αφορμή την παρατήρηση ότι το συναίσθημα επηρεάζει τα φωνήματα με διαφορετικό τρόπο, έχει προταθεί η δημιουργία νευρωνικού δικτύου για την κατηγοριοποίηση του άγχους που λαμβάνει υπόψη όχι μόνο τα χαρακτηριστικά ομιλίας για το άγχος αλλά και το φώνημα [95]. Πιο συγκεκριμένα, κάθε πρόταση διαιρείται σε frames και σε κάθε frame εξάγονται χαρακτηριστικά διέγερσης, άρθρωσης και φασματικά χαρακτηριστικά, καθώς και το είδος φωνήματος που εκφράζει. Ανάλογα με αυτά υπολογίζεται η πιθανοφάνεια κάθε πρότασης να ανήκει σε ένα είδος άγχους ως το άθροισμα των πιθανοφανειών για όλα τα frames και όλα τα φωνήματα [90]:

$$P(u|\Omega_c) = \sum_{q=1}^{Q} \sum_{\lambda=1}^{\Lambda} P(x_q|\Omega_c, \Theta_\lambda) P(\Theta_\lambda)$$

όπου u η πρόταση που μας ενδιαφέρει, x_q το $q_{\text{οστό}}$ frame Ω_c η κατάσταση άγχους c και Θ_λ το είδος φωνήματος λ.

Support Vector Machines (SVMs)

Τα τελευταία χρόνια έχει παρατηρηθεί μεγάλο ενδιαφέρον στην ταξινόμηση με SVMs, λόγω της μεγάλης τους ικανότητας γενίκευσης. Παρόλα αυτά τα SVMs είναι κατ΄ αρχάς δυαδικοί ταξινομητές και έχουν τον περιορισμό να δουλεύουν με διανύσματα σταθερού μήκους. Έχουν βρεθεί διάφορες τεχνικές που ξεπερνούν τα προβλήματα αυτά και καθιστούν δυνατή την αναγνώριση συναισθήματος μέσω SVMs.

Για την αντιμετώπιση του προβλήματος σταθερού μήκους διανύσματος και παράλληλα τη διατήρηση της χρήσιμης πληροφορίας, μπορεί να βρεθεί μία κατάλληλη συνάρτηση πυρήνα, που αντιστοιχεί τα μεταβαλλόμενου μήκους δεδομένα σε ένα χώρο σταθερής διάστασης. Προτάθηκε η συνάρτηση πυρήνα Fisher που μεταφέρει τα χαρακτηριστικά των παραθύρων μίας πρότασης σε ένα μονοδιάστατο χώρο [78]. Για μία πρόταση εισόδου x η συνάρτηση αυτή ορίζεται ως:

$$\mathbf{U}_x = \nabla_\theta log P(x|\theta)$$

όπου θ είναι το σύνολο των παραμέτρων ενός μοντέλου, π.χ. ενός GMM στο οποίο το θ είναι ο μέσος όρος, η τυπική απόκλιση και τα βάρη των γκαουσιανών. Για την ταξινόμηση περισσότερων από 2 κατηγορίες χρησιμοποιήθηκε μία μέθοδος που κατηγοριοποιεί ανά δύο τις κλάσεις με τη βοήθεια ασαφούς λογικής [83]. Ο μέσος όρος ακρίβειας για την κατηγοριοποίηση των 6 βασικών συναισθημάτων με τη μεθοδο αυτή είναι 97.65%, μέγεθος που αποτελεί και από τα υψηλότερα ποσοστά επιτυχίας στη βιβλιογραφία αναγνώρισης συναισθήματος.

Μία εναλλακτική μέθοδος για την κατηγοριοποίηση πολλών ομάδων είναι τα SVMs πολλαπλών επιπέδων [75]. Σε κάθε επίπεδο διαχωρίζεται μία κλάση συναισθημάτων από μία ομάδα κλάσεων, ώσπου να διαχωριστούν όλες οι κλάσεις. Το λάθος ταξινόμησης για τα 6 βασικά συναισθήματα και το ουδέτερο είναι 18.71%, χρησιμοποιώντας διάνυσμα διάστασης 33 με χαρακτηριστικά pitch, ενέργειας, διάρκειας και φάσματος (που έχουν διαλεχθεί με

τη μέθοδο LDA). Το λάθος αυτό είναι πολύ μικρότερο σε σχέση με το λάθος άλλων ταξινομητών που βασίζονται στο ίδιο διάνυσμα χαρακτηριστικών, όπως K-means, GMMs και νευρωνικά δίκτυα.

Ασαφή Σύνολα

Η ασαφής λογική χρησιμοποιείται στην αναγνώριση συναισθημάτος, γιατί έχει το πλεονέκτημα να μην απαιτεί πολλά δείγματα. Η ταξινόμηση των 6 βασικών συναισθημάτων (χαρά, λύπη, φόβος, θυμός, έκπληξη και αποστροφή) μπορεί να γίνει με διάνυσμα 18 χαρακτηριστικών: μέση ενέργεια, 14 LPC συντελεστές, διάρκεια, pitch και jitter, με τεχνικές ασαφούς λογικής [66]. Το κύριο μέλημα είναι η επιλογή της συνάρτησης συμμετοχής. Αν f_{ij} η τιμή του i χαρακτηριστικού για το j δείγμα εκπαίδευσης και N_i το πλήθος των δειγμάτων για το i χαρακτηριστικό, τότε για όλα τα χαρακτηριστικά $i = 1, \ldots, 18$ υπολογίζεται ο μέσος όρος και η τυπική απόκλιση των δειγμάτων εκπαίδευσης για κάθε συναίσθημα $r = 0, \ldots, 5$:

$$m_i(r) = \frac{1}{N_i}\sum_{j=1}^{N_i} f_{ij} \quad \text{και} \quad \sigma_i^2(r) = \frac{1}{N_i}\sum_{j=1}^{N_i}(f_{ij} - m_i)^2$$

Επειδή η έκφραση των συναισθημάτων μπορεί να γίνει με ποικίλους τρόπους, επιλέγεται η ακόλουθη συνάρτηση συμμετοχής του δείγματος $x = [x_1, \ldots, x_{18}]$ στο συναίσθημα r για κάθε χαρακτηριστικό i, στην οποία ο μέσος όρος και η τυπική απόκλιση έχουν διαφορετική βαρυτητα:

$$\mu(x_i, r) = exp(-\frac{(1-s) + s^2|x_i - m_i|}{(1+t) + t^2\sigma_i^2})$$

όπου s, t μεταβαλλόμενες παράμετροι ανάλογα με το χαρακτηριστικό για κάθε συναίσθημα. Οπότε η μέση συμμετοχή του δείγματος $x = [x_1, \ldots, x_{18}]$ στο συναίσθημα r είναι:

$$\mu_{average}(x, r) = \frac{1}{c}\sum_{i=1}^{18} \mu(x_i, r)$$

Χρησιμοποιώντας την παραπάνω σχέση με μεταβαλλόμενες τιμές των s, t για κάθε συναίσθημα, το νευρωνικό δίκτυο επιτυγχάνει 60,28% ποσοστό αναγνώρισης, ενώ με την κλασσική συνάρτηση συμμετοχής (στην οποία δε λαμβάνεται υπόψη η βαρύτητα των χαρακτηριστικών), το ποσοστό είναι σχεδόν το μισό (31.94%).

Δέντρα Απόφασης

Τα δέντρα απόφασης είναι μία αναπαράσταση κανόνων if-then-else. Το πλεονέκτημά τους είναι ότι τα δείγματα με τα οποία εκπαιδεύονται μπορεί να κωδικοποιούνται από σύβολα ή αριθμούς, καθώς και να είναι ατελή ή θορυβώδη.

Αρχικά, έχει χρησιμοποιηθεί δυαδικό δέντρο απόφασης [13], σε καθε κόμβο του οποίου αναγνωρίζεται ένα συναίσθημα, έτσι ώστε το δέντρο να έχει φύλλα σε όλα του τα επίπεδα (εκτός από τη ρίζα). Η απόφαση που παίρνεται σε κάθε κόμβο είναι ένα πρόβλημα δυαδικής ταξινόμησης, που βασίζεται σε μία τριπλέτα χαρακτηριστικών. Οι συγκεκριμένοι ερευνητές χωρίζουν τα χαρακτηριστικά αναγνώρισης συναισθήματος σε 3 ομάδες: χαρακτηριστικά συχνότητας (π.χ. pitch), χαρακτηριστικά ενέργειας και χαρακτηριστικά ρυθμού (π.χ. διάρκεια

εκφοράς ή διάρκεια παύσεων). Σε κάθε κόμβο του δέντρου χρησιμοποιείται μία τριπλέτα που περιέχει ένα χαρακτηριστικό από κάθε είδος. Στόχος είναι η εύρεση της βέλτιστης δομής του δέντρου και των πιο αποτελεσματικών τριπλέτων σε κάθε κόμβο. Επειδή τα συναισθήματα προς ταξινόμηση είναι λίγα και το διάνυσμα χαρακτηριστικών έχει μικρή διάσταση, μπορεί να γίνει εξαντλητική αναζήτηση των συνδυασμών, ώστε να προκύψει το βέλτιστο δυαδικό δέντρο. Τα αποτελέσματα αναγνώρισης του βέλτιστου δέντρου είναι 72.04% για τη βάση "Berlin Database of Emotional Speech" και τα χαρακτηριστικά που εμφανίστηκαν συχνότερα στα δέντρα αυτά ήταν το μέσο pitch, η μέγιστη τιμή του σήματος, η τυπική απόκλιση της ενέργειας, η κανονικοποιημένη διάρκεια και η μέση τιμή των τοπικών ελαχίστων του pitch.

Σε μία άλλη έρευνα [77] έχει προταθεί ο συνδυασμός συντακτικής και στατιστικής μάθησης για την κατηγοριοποίηση των 6 συναισθημάτων (θυμός, πλήξη, απαρέσκεια, χαρά, φόβος, λύπη) και του ουδέτερου στη βάση "Berlin Database of Emotional Speech". Στη συντακτική μάθηση τα δείγματα παρουσιάζονται ως μία δομή προτύπων, στην οποία λαμβάνονται υπόψη πιο περίπλοκες σχέσεις μεταξύ γνωρισμάτων σε σύγκριση με τη στατιστική μαθηση. Στο πρώτο μέρος αναγνώρισης (συντακτική μάθηση) τα γνωρίσματα του 39% δειγμάτων εκπαίδευσης κωδικοποιούνται με δομή δέντρου και εφαρμόζεται μία δεντροειδής γραμματική (tree grammar inference) για την εκπαίδευση 7 αυτόματων (ένα για καθε είδος συναισθήματος). Στο δεύτερο μέρος (στατιστική μάθηση) υπολογίζονται οι αποστάσεις όλων των δειγμάτων εκπαίδευσης από τα 7 αυτόματα που έχουν εξαχθεί και εφαρμόζεται ο αλγόριθμος C4.5 για τη δημιουργία δέντρου απόφασης. Η μέθοδος αυτή έχει ονομαστεί TGI+ (Tree Grammar Inference), όπου το '+' αντιπροσωπεύει την προσθήκη στατιστικής μάθησης. Στην έρευνα συγκρίνεται το αποτέλεσμα των νευρωνικών δικτύων με τα δέντρα απόφασης C4.5 και το TGI+ με μέσους όρους επιτυχίας 73.9%, 52.9% και 78.58% αντίστοιχα.

Συνδυασμός Τεχνικών

Συνδυασμός των παραπάνω τεχνικών κατηγοριοποίησης μπορεί να χρησιμοποιηθεί για την αποδοτικότερη ταξινόμηση των συναισθημάτων.

Ταξινομητές GMMs και k-κοντινότερων γειτόνων έχουν συνενωθεί με τη βοήθεια μιας απλής συνάρτησης ταξινόμησης. Στόχος της συγκεκριμένης έρευνας [45] ήταν η δυαδική κατηγοριοποίηση του θυμού από το ουδέτερο συναίσθημα. Για το λόγο αυτό χρησιμοποιήθηκαν GMMs με βάση MFCC χαρακτηριστικά, που συμβολίζονται ως M_1 και ταξινόμηση με k-κοντινότερους γείτονες με βάση το pitch, που είναι το M_2. Θεωρούμε ότι S_m για $m = 0, 1$ είναι οι πιθανοφάνειες η πρόταση s να ανήκει στην κλάση του θυμού και του ουδέτερου αντίστοιχα. Η πιθανοφάνεια που προκύπτει από τον GMM ταξινομητή για ένα σήμα s να ανήκει στην κλάση m συμβολίζεται με $LL(s|\Lambda_m^{M_1}) = log(P(s|\Lambda_m^{M_1}))$, όπου $\Lambda_m^{M_1}$ οι παράμετροι του GMM για την κλάση m με βάση τα MFCC χαρακτηριστικά M_1. Επειδή η πιθανοφάνεια του GMM έχει πολύ μικρότερη τιμή από αυτή των k-κοντινότερων γειτόνων, η πιθανοφάνεια των τελευταίων δε λογαριθμίζεται και συμβολίζεται ως $P(s|\Lambda_m^{M_2})$, όπου $\Lambda_m^{M_2}$ οι παράμετροι του ταξινομητή για την κλάση m με βάση το pitch M_2. Η συνάρτηση συνένωσης των δύο ταξινομητών είναι:

$$C(s) = \left\{ \begin{array}{ll} H_0, & \xi \leq \theta \\ H_1, & \xi < \theta \end{array} \right\}$$

για $\xi = S_0 - S_1 = \lambda \cdot P(s|\Lambda_0^{M_2}) - \lambda \cdot P(s|\Lambda_1^{M_1}) + (1-\lambda) \cdot log(P(s|\Lambda_0^{M_2}) - (1-\lambda) \cdot log(P(s|\Lambda_1^{M_2})$

και θ ένα κατώφλι.
Τα αποτελέσματα που προέκυψαν με βάση τη συνάρτηση συνένωσης είχαν ελαφρώς μικρότερα σφάλματα σε σχέση με αυτά που προήλθαν από την ανεξάρτητη λειτουργία των δύο ταξινομητών.

Ένας απλός συνδυασμός GMMs και HMMs έχει γίνει για την ταξινόμηση 5 συναισθημάτων (θυμός, φόβος, χαρά, λύπη, έκπληξη) και του ουδέτερου [40]. Η εκπαίδευση και ο έλεγχος των GMMs και HMMs γίνεται με βάση χαρακτηριστικά θεμελιώδης συχνότητας, ενέργειας και διάρκειας, οπότε προκύπτουν οι παράμετροι G_i και H_i για $i = 1, \ldots, M$, όπου M το πλήθος των κλάσεων, για τα δύο μοντέλα αντίστοιχα. Στη συνέχεια κάθε σήμα εισόδου S_j με χαρακτηριστικά X_j για το GMM και Y_j για το HMM ελέγχεται από τους δύο ταξινομητές. Έτσι εξάγονται οι πιθανοφάνειες $Z_j(i) = log(P(X_j|G_i)$ και $Z_j(M+i) = log(P(Y_j|H_i)$, που εκφράζουν την πιθανότητα το σήμα S_j να ανήκει στην κλάση i σύμφωνά με το GMM και HMM μοντέλο. Ο συνδυασμός των δύο μοντέλων μπορεί να γίνει με ένα σταθμισμένο μπεϋζιανό ταξινομητή σύμφωνα με την παρακάτω σχέση:

$$r_j = arg \max_{0<i<\leq M} [(1-\alpha) \cdot log(P(X_j|G_i)) + \alpha \cdot log(P(Y_j|H_i))]$$

Το καλύτερο αποτέλεσμα επιτυγχάνεται για $\alpha = 0.8$ και είναι 80%. Επίσης οι πιθανοφάνειες των GMM και HMM ταξινομητών μπορεί να συνδυαστούν με ένα νευρωνικό δίκτυο ενός κρυφού στρώματος με 50 κόμβους, οπότε προκύπτει 83.1% ποσοστό επιτυχίας.

Η βελτίωση των αποτελεσμάτων ασταθών ταξινομητών, όπως τα νευρωνικά δίκτυα ή τα δέντρα αποφάσεων, μπορεί να επιτευχθεί παράγοντας πολλούς τέτοιους ταξινομητές και συνδυάζοντάς τους σε ένα σύνολο που ονομάζεται "ensemble". Δύο από τις πιο γνωστές μεθόδους για αυτό είναι το Bagging και το Boosting. Έστω C_t για $t = 1, \ldots, T$ οι αρχικοί ταξινομητές και C^* ο τελικός ταξινομητής. Στην τεχνική του Bagging κάθε ταξινομητής C_t κατηγοριοποιεί όλα τα δείγματα και μία απλή διαδικασία πλειοψηφίας ακολουθείται για την τελική απόφαση του ταξινομητή C^*. Στο Boosting τα δείγματα για κάθε ταξινομητή C_t δίνονται τυχαία και ο τελικός ταξινομητής C^* αποφασίζει επίσης με τη μέθοδο της πλειοψηφίας δίνοντας όμως τα κατάλληλα βάρη σε κάθε ταξινομητή. Σε μία έρευνα [72] αποδείχθηκε ότι ο συνδυασμός ταξινομητών δίνει καλύτερα αποτελέσματα για την κατηγοριοποίηση 6 συναισθημάτων (χαρά, θυμός, αποστροφή, φόβος, λύπη, έκπληξη) και του ουδέτερου. Με βάση διάνυσμα διάστασης 100, που περιέχει χαρακτηριστικά pitch, ενέργειας, διάρκειας και φάσματος, το $C4.5$ δέντρο απόφασης επιτυγχάνει 44.59% σωστή αναγνώριση. Ο συνδυασμός 4 τέτοιων δέντρων με Bagging βελτιώνει το αποτέλεσμα σε 54.04%, ενώ με Boosting σε 67.56%. Η τεχνική του Boosting χρησιμοποιείται και σε μία ακόμα έρευνα, στην οποία χρησιμοποιούνταν MFCC και ESA χαρακτηριστικά και η κατηγοριοποίηση γίνεται με GMMs [68]. Το αποτέλεσμα του συνδυασμού αυτού για την αναγνώριση του άγχους είναι 79.35% και επιφέρει βελτίωση +12.99% σε σύγκριση με το μεμονωμένο GMM ταξινομητή.

Διαφορετικών ειδών ταξινομητές μπορούν να συνδυαστούν για καλύτερα αποτελέσματα. Στην έρευνα που αναφέραμε προηγουμένως [72], μία από τις μεθόδους που χρησιμοποιούνται είναι η StackingC, με την οποία συνδυάζονται ένα SVM, ένα νευρωνικό δίκτυο, ένα δέντρο απόφασης $C4.5$ και ένας κατηγοριοποιητής k-κοντινότερων γειτόνων με αποτέλεσμα 71.62%.

Κεφάλαιο 4

Μελέτη Χαρακτηριστικών Διάρκειας, Προσωδίας και Φωνητικού Σωλήνα

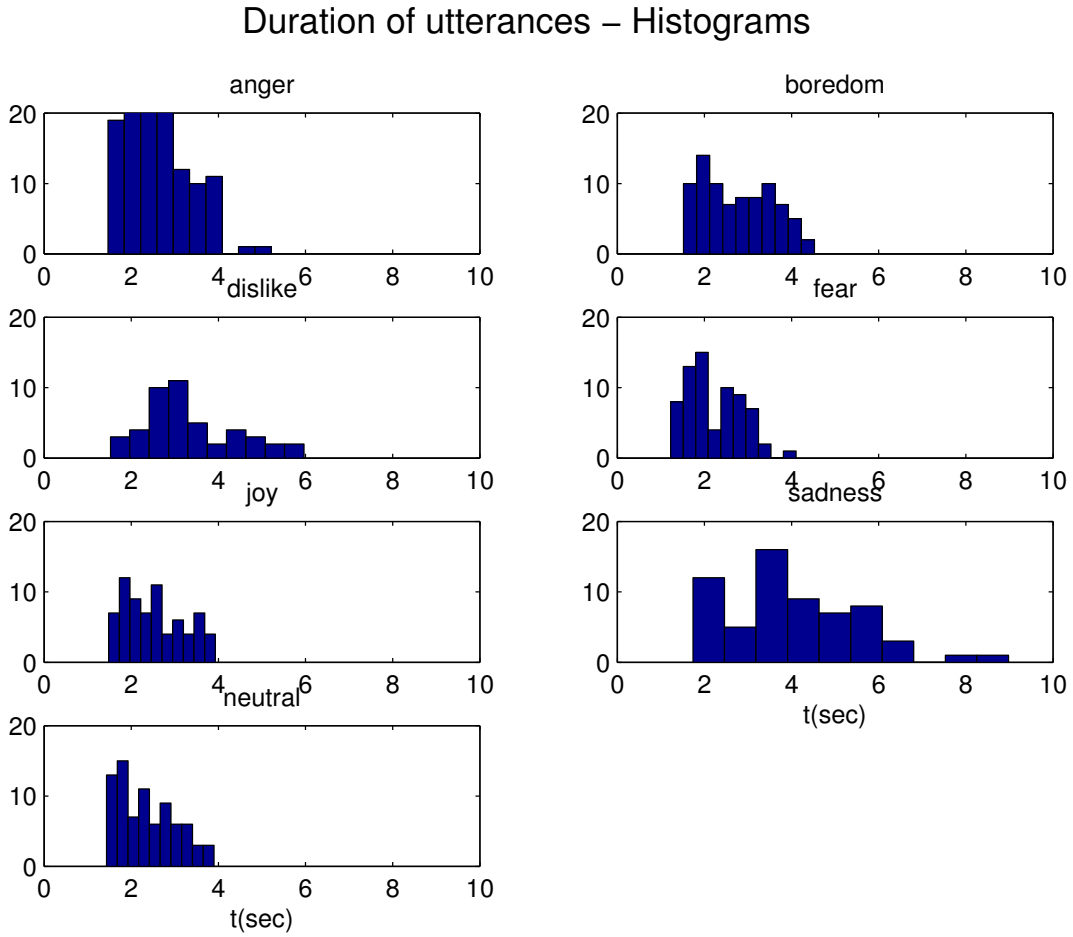
Στο κεφάλαιο αυτό εξετάζουμε τα χαρακτηριστικά που έχουν χρησιμοποιηθεί ευρέως στην αναγνώριση συναισθήματος και είναι η διάρκειας εκφοράς, η θεμελιώδης συχνότητα και οι διαμορφώτριες συχνότητες. Τα χαρακτηριστικά αυτά αντικατοπτρίζουν τη μελωδικότητα του λόγου και τη μορφή του φωνητικού σωλήνα. Επίσης, μελετάμε τους χαμηλόσυχνους συντελεστές του μετασχηματισμού Fourier, καθώς είναι ένα απλό χαρακτηριστικό που υπολογίζεται εύκολα και απεικονίζει τις χαμηλόσυχνες διακυμάνσεις της φωνής.

4.1 Χρονική Διάρκεια Εκφοράς

Εξετάζουμε τη χρονική διάρκεια εκφοράς για όλα τα συναισθήματα. Στο σχήμα 4.1 φαίνονται ιστογράμματα για κάθε συναίσθημα ξεχωριστά του μεγέθους αυτού για όλες τις προτάσεις και όλους τους ομιλητές. Παρατηρούμε ότι μεγαλύτερη διάρκεια έχουν προτάσεις που εκφράζονται με λύπη, ενώ σύντομες είναι οι προτάσεις χαράς, φόβου και πλήξης. Με ενδιάμεση διάρκεια εκφοράς εκφράζονται ο θυμός και η απαρέσκεια.

Προσπαθώντας να μηδενίζουμε τον παράγοντα εξάρτησης από το περιεχόμενο της πρότασης, στο σχήμα 4.2 απεικονίζουμε τη διάρκεια εκφοράς κάθε πρότασης για κάθε συναίσθημα. Οι παραπάνω παρατηρήσεις επιβεβαιώνονται και σε αυτή την περίπτωση. Η οριζόντια γραμμή που αντιστοιχεί στο συναίσθημα της λύπης αποτελείται από πράσινο, κίτρινο, κόκκινο και μπορντό χρώμα, που σημαίνει μεγάλες τιμές διάρκειας εκφοράς. Αντίθετα, στην οριζόντια γραμμή του φόβου βλέπουμε μόνο μπλε και γαλάζια χρώματα, που αντιστοιχούν σε μικρή διάρκεια εκφοράς.

Αν και η διάρκεια εκφοράς δεν είναι το πλέον διακριτό χαρακτηριστικό στην αναγνώριση συναισθήματος, από τα αποτελέσματα που είδαμε, μπορούμε να πούμε ότι η λύπη συνήθως εκφράζεται με αργό λόγο, ενώ η χαρά με πιο κοφτό λόγο.



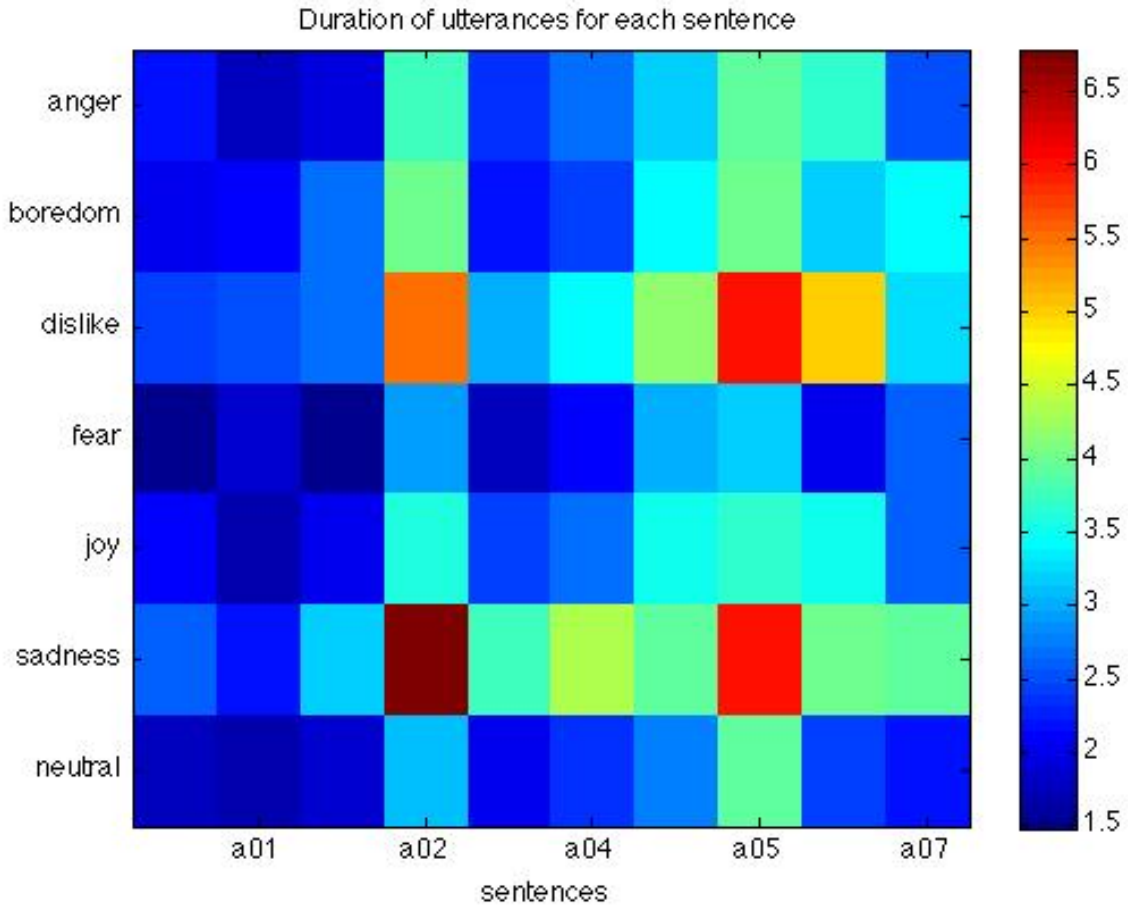
Σχήμα 4.1: Ιστόγραμμα διάρκειας εκφοράς για όλες τις προτάσεις σε κάθε συναίσθημα

4.2 Θεμελιώδης Συχνότητα (Pitch)

Η θεμελιώδης συχνότητα είναι το πιο συνηθισμένο χαρακτηριστικό που χρησιμοποιείται στην αναγνώριση συναισθήματος. Από τον ορισμό η θεμελιώδης συχνότητα βρίσκεται από τη θέση μεγίστου της αυτοσυσχέτισης του σήματος φωνής, ενώ ο βαθμός της περιοδικότητάς του σήματος από το σχετικό ύψος του μεγίστου αυτού. Ο αλγόριθμος που ακολουθούμε για τον εντοπισμό της θεμελιώδους συχνότητας [5] συνοψίζεται στα παρακάτω βήματα.

1. **Χρήση παραθύρου Hanning**
   Χρήση του παραθύρου Hanning $w(t) = \frac{1}{2} - \frac{1}{2}cos(\frac{2\pi t}{T})$ για πολλαπλασιασμό με το αρχικό σήμα. Απομόνωση του πλευρικού λοβού του παραθύρου Hanning φιλτράροντας στο χώρο των συχνοτήτων από το 95% έως το 100% της συχνότητας Nyquist. Υπολογισμός της αυτοσυσχέτισης του παραθύρου Hanning.

$$r_w(t) = (1 - \frac{|\tau|}{T})(\frac{2}{3} + \frac{1}{3}cos\frac{2\pi\tau}{T}) + \frac{1}{2\pi}sin\frac{2\pi|\tau|}{T}$$



Σχήμα 4.2: Διάρκεια εκφοράς κάθε προτάσης σε κάθε συναίσθημα

Το παράθυρο Hanning είναι πιο στενό από το ορθογώνιο, το Hamming και το Welch και για το λόγο αυτό είναι και πιο ανθεκτικό σε γρήγορα μεταβαλλόμενα σήματα.

2. **Χωρισμός του σήματος φωνής σε frames**
   Ανάλυση του αρχικού σήματος φωνής $x(t)$ σε frames 400 δειγμάτων (25ms) με επικάλυψη 120 δείγματα (7.5ms) Για κάθε frame αναζητάμε το πολύ *MaximumNumberOfCandidatesPerFrame* υποψήφια σημεία περιοδικότητας. Ο αριθμός αυτός συμπεριλαμβάνει και το άφωνο υποψήφιο σημείο που είναι πάντα παρόν.

3. **Υπολογισμός της αυτοσυσχέτισης για κάθε παραθυροποιημένο frame**
   Έστω ότι παίρνουμε frame διάρκειας $T = 25ms$ με κέντρο γύρω από το σημείο t_{mid}, για το οποίο γίνονται τα παρακάτω βήματα.

   (α′) Αφαίρεση του μέσου όρου από τις τιμές του frame και πολλαπλασιασμός με το παράθυρο Hanning $w(t)$.

$$s(t) = (x(t_{mid} - \frac{1}{T}T + t) - \mu_x)w(t)$$

(b) Υπολογισμός της κανονικοποιημένης αυτοσυσχέτισης του frame.

$$r_s(\tau) = r_s(-\tau) = \frac{\int_0^{T-\tau} s(t)s(t+\tau)dt}{\int_0^T s^2(t)dt}$$

Στην υλοποίησή μας ο υπολογισμός της αυτοσυσχέτισης γίνεται μέσω του Fast Fourier μετασχηματισμού. Αρχικά, μεταφέρεται το σήμα στο χώρο των συχνοτήτων

$$\tilde{s}(\omega) = \int s(t) \exp^{-i\omega t} dt$$

και στη συνέχεια, υπολογίζεται ο αντίστροφος μετασχηματισμός Fourier της πυκνότητας της ενέργειας του σήματος $|\tilde{s}(\omega)^2|$, που μεταφέρει το σήμα στο χώρο της καθυστέρησης (lag).

$$r_s(\tau) = \int |\tilde{s}(\omega)|^2 \exp^{i\omega\tau} \frac{d\omega}{2\pi}$$

(g) Υπολογισμός της αυτοσυσχέτισης του παραθυροποιημένου frame.

$$r_x(\tau) = \frac{r_s(\tau)}{r_w(\tau)}$$

4. **Εύρεση των υποψήφιων θέσεων της θεμελιώδους συχνότητας**
Εφόσον το αρχικό σήμα έχει δειγματοληπτθεί με περίοδο δειγματοληψίας $\Delta\tau$, η αυτοσυσχέτιση για κάθε frame είναι σήμα διακριτού χρόνου r_n. Για να βρούμε ένα τοπικό μέγιστο στο διάστημα $(m-1)\Delta\tau$ έως $(m+1)\Delta\tau$ αρκεί να ισχύει

$$r_m > r_{m-1}$$

Έτσι, μία πρώτη εκτίμηση της περιόδου του pitch θα ήταν $\tau_{max} \approx m\Delta\tau$, η οποία όμως δεν είναι πολύ ακριβής. Για το λόγο αυτό εφαρμόζουμε παραβολική παρεμβολή γύρω από το $m\Delta\tau$ χρησιμοποιώντας N δείγματα από αριστερά και δεξιά. Με τη βοήθεια ενός παραθύρου Hanning για να μειώσουμε την επίδραση των απότομων μεταβολών, η συνάρτηση αυτοσυσχέτισης γίνεται:

$$r(\tau) = \sum_{n=1}^{N} r_{n_r-n} \frac{sin\pi(\phi_l+n-1)}{\pi(\phi_l+n-1)}(\frac{1}{2}+\frac{1}{2}cos\frac{\pi(\phi_l+n-1)}{\phi_l+N})+$$

$$\sum_{n=1}^{N} r_{n_r+n} \frac{sin\pi(\phi_l+n-1)}{\pi(\phi_l+n-1)}(\frac{1}{2}+\frac{1}{2}cos\frac{\pi(\phi_l+n-1)}{\phi_l+N})$$

όπου $n_l \equiv$ μεγαλύτερος ακέραιος $\leq \frac{\tau}{\Delta\tau}$, $n_r \equiv n_l+1$, $\phi_l \equiv \frac{\tau}{\Delta\tau} - n_l$ και $\phi_r \equiv 1-\phi_l$. Τα σημεία και οι τιμές τοπικών μεγίστων στην προηγούμενη εξίσωση μπορούν να εντοπιστούν με μεγάλη ακρίβεια και είναι αυτά που αντιστοιχούν σε θεμελιώδη συχνότητα ανάμεσα στις τιμές *MinimumPitch* και *MaximumPitch*. Το πρώτο υποψήφιο άφωνο σημείο για τη θέση του pitch έχει τοπική ένταση

$$R \equiv VoicingThreshold + \max(0.2 - \frac{LocalAbsolutePeak/GlobalAbsolutePitch}{SilenceThreshold/(1+VoicingThreshold)})$$

Τα υποψήφια έμφωνα σημεία έχουν τοπική ένταση

$$R \equiv (\tau_{max}) - OctaveCost^2 \cdot log(MinimumPitch \cdot \tau_{max})$$

Με τον τρόπο αυτό επιλέγονται ζευγάρια συχνότητας - έντασης (F_{ni}, R_{ni}), όπου ο δείκτης n απεικονίζει τα frames και το i τα υποψήφια σημεία του pitch σε κάθε frame.

5. **Εντοπισμός καλύτερης υποψήφιας θεμελιώδους συχνότητας μέσω αλγορίθμου εύρεσης καθολικά βέλτιστου μονοπατιού**
Το καλύτερο υποψήφιο pitch σε κάθε frame n είναι αυτό με τη μεγαλύτερη τιμή έντασης R. Με την προσέγγιση αυτή όμως μπορεί να υπάρξουν πάνω από ένα βέλτιστα σημεία με παραπλήσιες τιμές έντασης και για το λόγο αυτό χρησιμοποιείται ένας αλγόριθμος εύρεσης καθολικά βέλτιστου μονοπατιού. Έστω p_n ένας αριθμός μεταξύ 1 και του πλήθους των υποψήφιων θέσεων pitch για κάθε frame n. Οι τιμές $\{p_n | 1 \leq n \leq NumberOfFrames\}$ ορίζουν ένα μονοπάτι με τις υποψήφιες θέσεις pitch και κάθε μονοπάτι σχετίζεται με ένα κόστος

$$cost(\{p_n\}) = \sum_{n=2}^{NumberOfFrames} transitionCost(F_{n-1,p_{n-1}} F_{n,p_n}) - \sum_{n-1}^{NumberOfFrames} R_{np_n}$$

όπου η συνάρτηση του κόστους μετάβασης ορίζεται ως

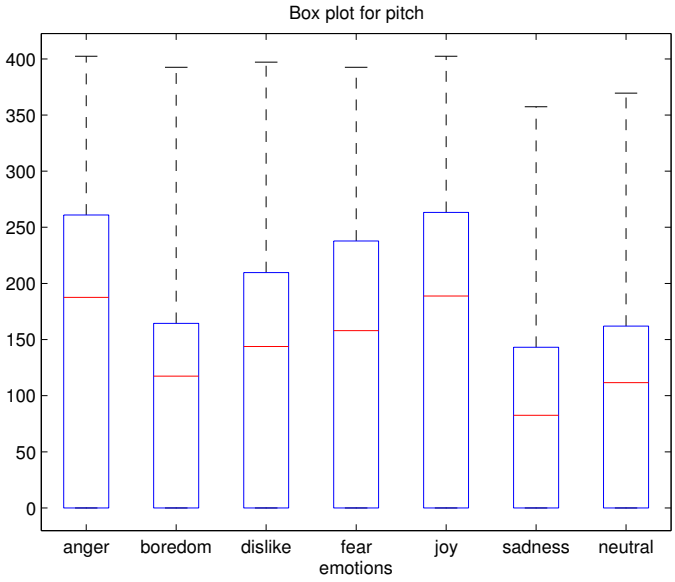
$$TransitionCost(F_1, F_2) = \left\{ \begin{array}{c} 0, \text{ if } F_1 = 0 \text{ and } F_2 = 0 \\ VoicedUnvoicedCost, \text{ if } F_1 = 0 \text{ xor } F_2 = 0 \\ OctaveJumpCost^2 \cdot |log\frac{F_1}{F_2}| \text{ if } F_1 \neq 0 \text{ and } F_2 \neq 0 \end{array} \right\}$$

Το καθολικά βέλτιστο μονοπάτι είναι αυτό με το λιγότερο κόστος και η θεμελιώδης συχνότητα εντοπίζεται γυρίζοντας στην αρχή του μονοπατιού αυτού.

Στο σχήμα 4.3 σχεδιάζουμε το box plot του για κάθε συναίσθημα. Σε κάθε ορθογώνιο σχήμα η κόκκινη γραμμή αντιστοιχεί στη διάμεσο των τιμών, ενώ η άνω και κάτω μπλε γραμμή στο 25% και 75% εκατοστιαίο σημείο.

Τα έντονα συναισθήματα σχετίζονται με υψηλή τιμή και διασπορά θεμελιώδους συχνότητας, ενώ τα λιγότερο έντονα συναισθήματα με χαμηλότερη τιμή και διασπορά [94]. Από τα box plots βλέπουμε ότι ο θυμός, η χαρά και ο φόβος παρουσιάζουν μεγάλες τιμές θεμελιώδους συχνότητας, ενώ η λύπη, η πλήξη και το ουδέτερο έχουν χαμηλές τιμές. Το εύρος τιμών της θεμελιώδους συχνότητας είναι μεγάλο για το θυμό και τη χαρά, σχετικά μεγάλο για το φόβο, και πιο μικρό για τη λύπη και την πλήξη. Οι παρατηρήσεις αυτές βρίσκονται σε συμφωνία και με προηγούμενες έρευνες [14, 96].

Στη συνέχεια, μελετάμε τη μεταβλητότητα της θεμελιώδους συχνότητας, υπολογίζοντας της διαφορά των τιμών μεταξύ διαδοχικών παραθύρων. Στο σχήμα 4.4 βρίσκεται η γραφική παράσταση του μεγέθους αυτού για μία συγκεκριμένη πρόταση με τρία διαφορετικά συναισθήματα και το ουδέτερο. Ο θυμός και η χαρά παρουσιάζουν πιο συχνές και απότομες μεταβολές της θεμελιώδους συχνότητας. Διαφοροποιούνται ως προς την εξέλιξη των τιμών της θεμελιώδους συχνότητας, που για το θυμό είναι φθίνουσα, ενώ για τη χαρά είναι σταθερή. Η λύπη, ο φόβος και το ουδέτερο έχουν πιο ομαλή γραφική παράσταση, ενώ η πλήξη έχει απότομες μεταβολές, όχι όμως τόσες όσες ο θυμός και η χαρά.



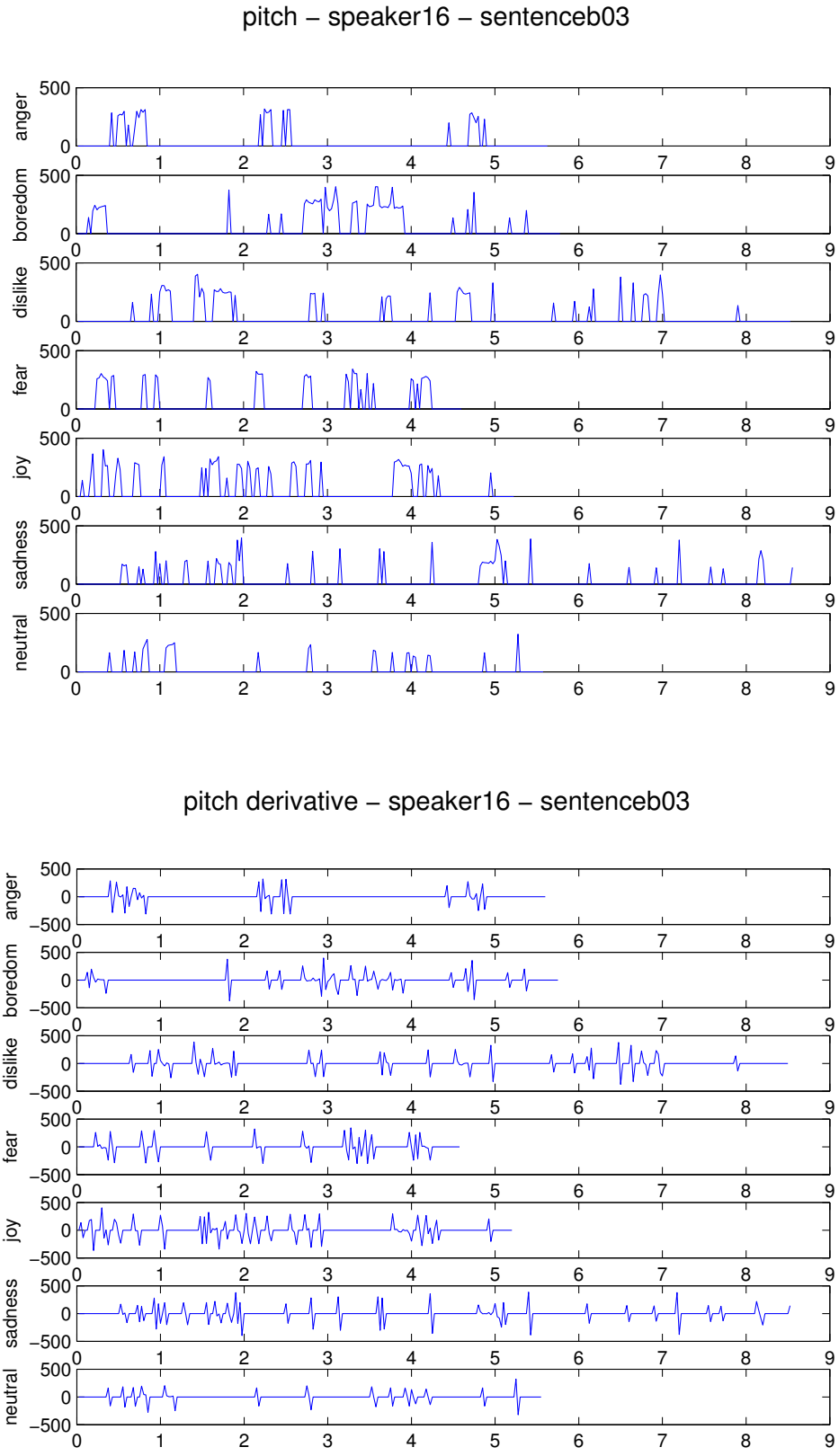
Σχήμα 4.3: Box plot θεμελιώδους συχνότητας κάθε πρότασης σε κάθε συναίσθημα. Σε κάθε box plot η κόκκινη γραμμή παριστάνει τη διάμεσο, η άνω και κάτω μπλε γραμμή του ορθογωνίου το 75% και 25% εκατοστιαίο σημείο και η άνω και κάτω μαύρη γραμμή τη μέγιστη και ελάχιστη τιμή.

Επειδή οι τιμές του pitch εξαρτώνται από το φύλο και τα φωνητικά χαρακτηριστικά του ομιλητή, μελετάμε τις μεταβολές του μεγέθους αυτού, που έχουν σχέση με την προσωδία του λόγου. Πιστεύεται ότι η προσωδία είναι αυτή που διαμορφώνει την έκφραση του συναισθήματος.

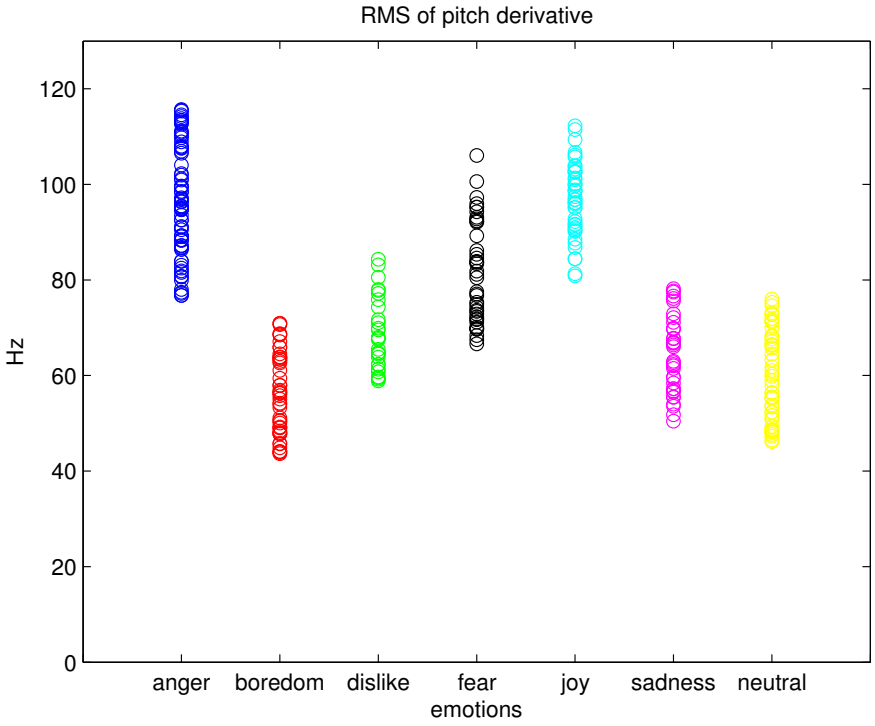
Για κάθε πρόταση βρίσκουμε την παράγωγο του pitch, υπολογίζοντας τη διαφορά μεταξύ δύο διαδοχικών δειγμάτων. Στη συνέχεια βρίσκουμε την τυπική απόκλιση του μεγέθους αυτού από τη σχέση

$$rms_{pitch'} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[pitch' - mean_{pitch'}]^2}$$

όπου $pitch' = \frac{d(pitch)}{dt}$ είναι η παράγωγος του pitch. Με τον τρόπο αυτό εξετάζουμε τη μεταβλητότητα του pitch για τα διάφορα συναισθήματα. Ένα διάγραμμα που απεικονίζει τις τιμές του μεγέθους αυτού φαίνεται στο σχήμα 4.5. Σημειώνουμε ότι στο σχήμα δεν απεικονίζουμε τους outliers, δηλαδή τις άνω και κάτω ακρότατες τιμές για κάθε συναίσθημα. Ο θυμός, η χαρά και ο φόβος εμφανίζουν τις μεγαλύτερες αποκλίσεις. Αυτό είναι αναμενόμενο, γιατί όπως έχουμε αναφέρει, τα συναισθήματα του θυμού και της χαράς είναι πιο έντονα, οπότε έχουν και πολλές μεταβολές. Οι μεγάλες τιμές της τυπικής απόκλισης για το φόβο μπορούν να δικαιολογηθούν από το γεγονός ότι στο συναίσθημα αυτό τρεμοπαίζει η φωνή, οπότε μεταβάλλεται έντονα και το pitch. Αντίθετα, η πλήξη, η λύπη και το ουδέτερο παρουσιάζουν μικρότερη μεταβολή της παραγώγου του pitch. Συνήθως στα συναισθήματα αυτά η φωνή είναι πιο σταθερή, οπότε δεν υπάρχουν και έντονες μεταβολές.

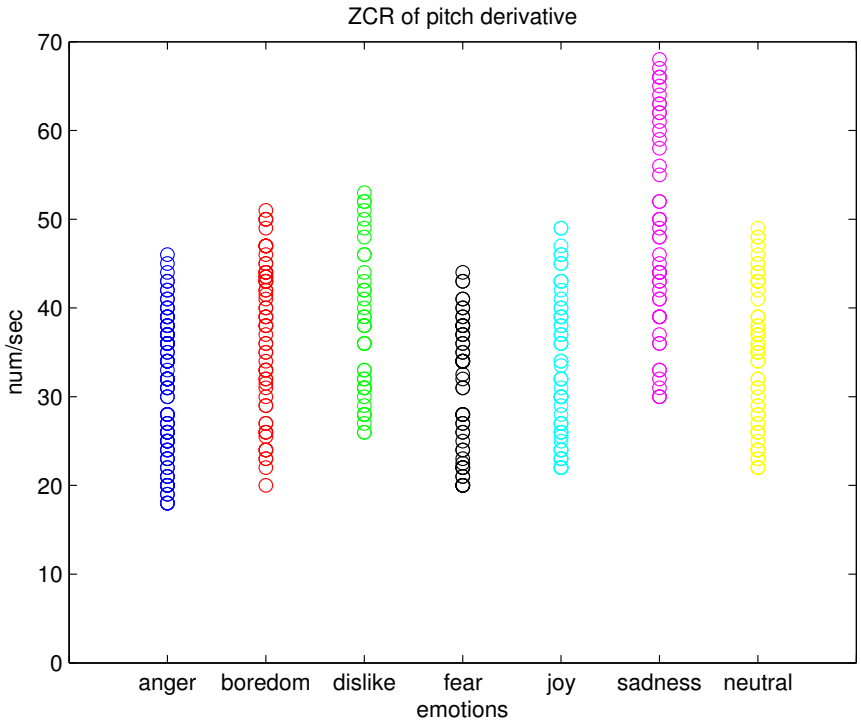


Σχήμα 4.4: Γραφική παράσταση της χρονικής εξέλιξης του pitch και της παραγώγου του.



Σχήμα 4.5: Διάγραμμα του εύρους των τιμών της τυπικής απόκλισης της παραγώγου του pitch για όλα τα συναισθήματα.

Από το σχήμα 4.4 της παραγώγου του pitch για την πρόταση b03, παρατηρούμε ότι διαφέρει ο αριθμός των zero-crossings για τις γραφικές παραστάσεις των διαφορετικών συναισθημάτων. Έτσι, στο σχήμα 4.6 παρουσιάζεται ένα διάγραμμα με το Zero-Crossing-Rate (ZCR) της παραγώγου του pitch, που μετριέται στις προτάσεις όλων των συναισθημάτων. Το συναίσθημα που διαφέρει πολύ σε σχέση με τα υπόλοιπα είναι η λύπη, για την οποία παρατηρούνται πολλές υψηλές τιμές ZCR της παράγωγο του pitch. Αυτό σημαίνει ότι η παράγωγος του pitch της λύπης έχει μικρότερη ενέργεια από τα υπόλοιπα συναισθήματα, όπως άλλωστε αναμένουμε και από τον χαμηλό τόνο εκφοράς του συναισθήματος αυτού.



Σχήμα 4.6: Διάγραμμα του εύρους των τιμών του ZCR της παραγώγου του pitch για όλα τα συναισθήματα.

4.3 Διαμορφώτριες Συχνότητες (Formants)

Ο όρος formant παραπέμπει στις κορυφές του φασματογράμματος ενός ήχου και σχετίζεται με τη διαμόρφωση της πηγής του ηχητικού σήματος. Η διάκριση των διαφόρων φωνηέντων είναι δυνατή μέσω των περιοχών των formants.

Ένας αξιόπιστος και απλός υπολογισμός των formants μπορεί να γίνει μέσω της Γραμμικής Πρόβλεψης (LPC Analysis) [50]. Βρίσκουμε το πολυώνυμο πρόβλεψης με συντελεστές a_n, $n = 1, \ldots, N$, όπου $N = 2 + \frac{Fs}{1000}$ και Fs συχνότητα δειγματοληψίας. Στη συνέχεια παραγοντοποιούμε το πολυώνυμο αυτό, ώστε να βρεθούν οι μέγιστοι το πλήθος $N/2$ συζυγείς μιγαδικοί πόλοι του. Υπολογίζουμε τις φάσεις των μιγαδικών αυτών πόλων και οι 4 μικρότερες γωνίες αντιστοιχούν στα 4 πρώτα formants. το πλεονέκτημα της μεθόδου αυτής είναι ότι προσφέρει αξιόπιστο τρόπο εύρεσης των formants και του εύρους αυτών. Επίσης, δε λαμβάνει υπόψη της μεγάλο πλήθος άσχετων πόλων, γιατί αυτοί χαρακτηρίζονται από υπερβολικά μεγάλο εύρος formants. Το μειονέκτημα της μεθόδου αυτής είναι ότι το μοντέλο των πόλων που χρησιμοποιείται δεν προσομοιάζει με επιτυχία τους ένρινους ήχους. Επίσης, ο τρόπος υπολογισμού του εύρους των formants δεν είναι πάντα ακριβής.

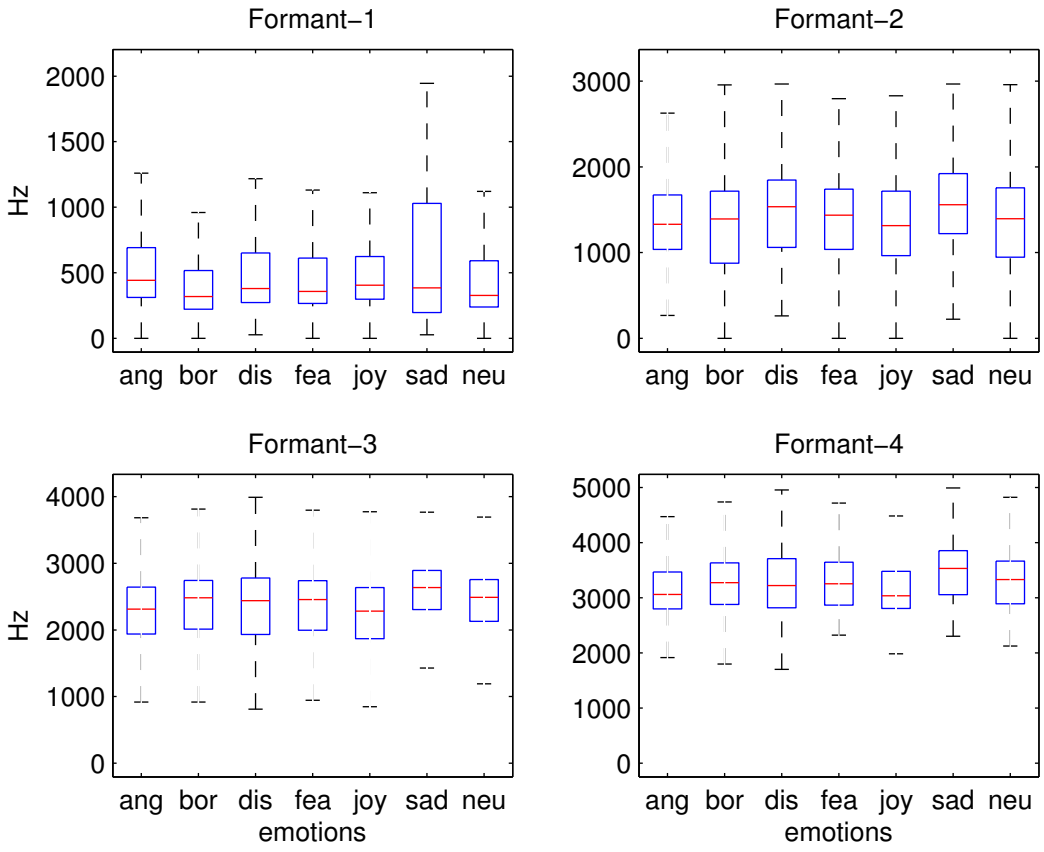
Στο σχήμα 4.7 φαίνονται τα 4 πρώτα formants για 6 συναισθήματα και το ουδέτερο. Το συναίσθημα που ξεχωρίζει όσον αφορά στο 1ο formant είναι η λύπη, που παρουσιάζει μεγαλύτερο εύρος τιμών και έχει πολλές τιμές στο διάστημα $1000 - 2000Hz$. Αυτό σημαίνει ότι υπάρχει ολίσθηση του 1ου formant σε υψηλότερες τιμές από $1000Hz$ για το συναίσθημα αυτό. Η λύπη έχει και στα formants 2-4 πιο αυξημένες τιμές από τα υπόλοιπα συναισθήματα. Γενικά, οι εμφανέστερες διαφορές μεταξύ των συναισθημάτων βρίσκονται στο 1ο και 2ο formant και λιγότερο στο 3ο και το 4ο.

Σε προηγούμενες έρευνες έχει υποστηριχθεί η υπόθεση ότι τα διαφορετικά συναισθήματα επηρεάζουν με ποικίλους τρόπους τις ιδιότητες των διαφόρων ειδών ήχου και πιο συγκεκριμένα τις ιδιότητες των φωνηέντων [48]. Τα φωνήεντα είναι ήχοι που προφέρονται με ανοιχτό φωνητικό σωληνα, δηλαδή δεν υπάρχει κάποιο κλείσιμο ή στένωση κατά την εκφορά τους. Αυτός είναι και ο λόγος που η συναισθηματική επένδυση της ομιλίας έχει μεγαλύτερη επίδραση στα φωνήεντα. Η βάση δεδομένων EMO-DB είναι στη γερμανική γλώσσα, οπότε εξετάζουμε 4 φωνήεντα που υπάρχουν σε αυτή: 'a', 'e', 'ie', 'ö'. Μέσω του Praat εντοπίζουμε τα φωνήεντα αυτά σε προτάσεις που υπάρχουν στη βάση δεδομένων και υπολογίζουμε τις 4 πρώτες διαμορφώτριες συχνότητες (formants). και τη θεμελιώδη συχνότητα (pitch). Στον πίνακα 4.1 δείχνουμε τις λέξεις από τις οποίες πήραμε τα 4 φωνήεντα.

Φώνημα	Λέξεις
a	λαπεν (τραπεζομάντηλο), sagen (λέω)
ö	κönντε (θα μπορούσα)
e	δερ (το)
ie	λιεγτ (κείται), sieben (επτά)

Πίνακας 4.1: 4 γερμανικά φωνήματα και οι λέξεις στις οποίες βρίσκονται.

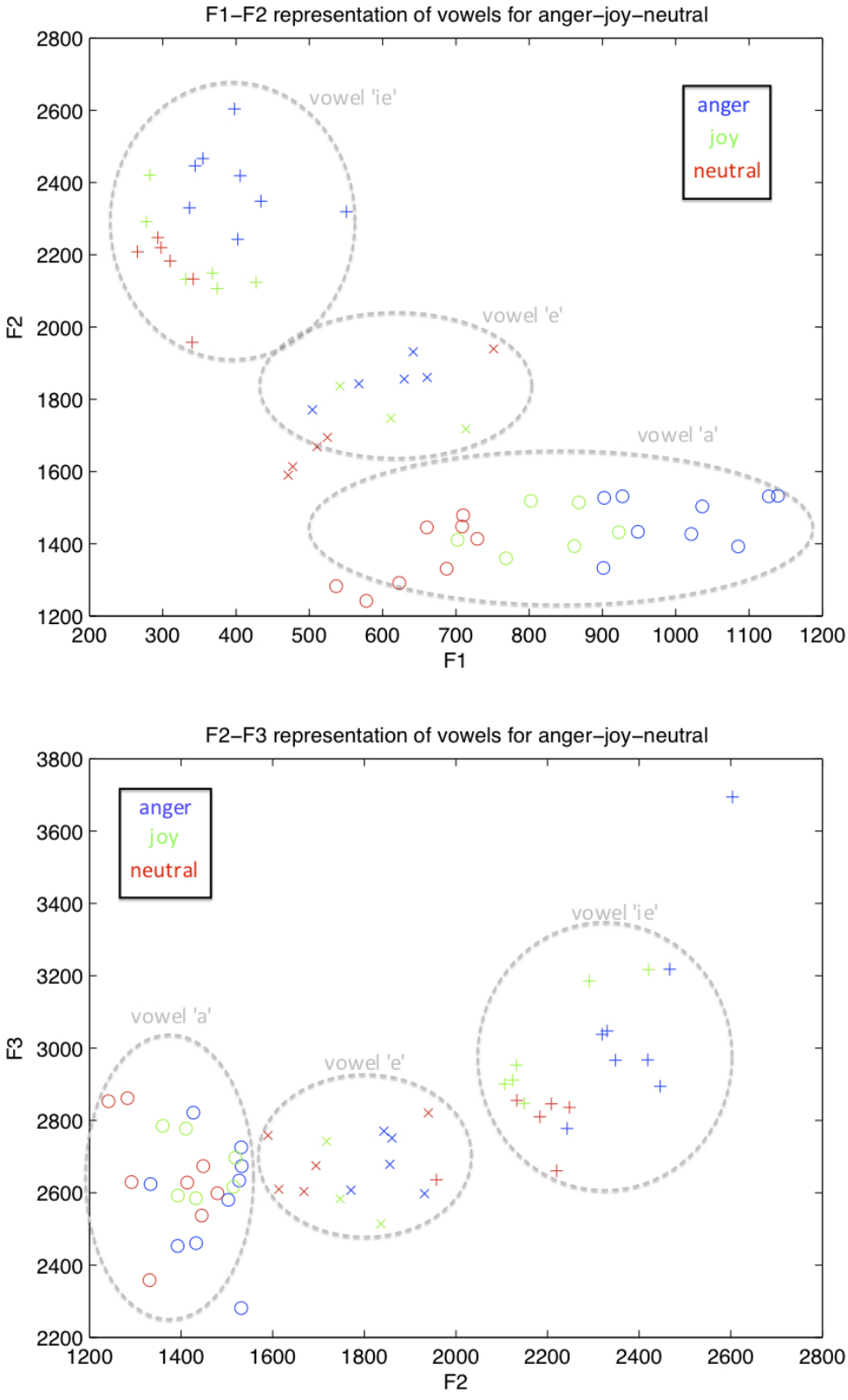
Στο σχήμα 4.8a φαίνεται η γραφική παράσταση των ζευγών των τιμων F1 και F2 για 3 φωνήεντα και 3 συναισθήματα. Παρατηρούμε ότι τα 3 συναισθήματα (και κυρίως ο θυμός και το ουδέτερο) διαχωρίζονται με μεγάλη διακριτότητα σε όλα τα φωνήεντα με βάση την F1-F2 απεικόνιση. Αντίθετα με βάση την F2-F3 απεικόνιση ο διαχωρισμός δεν είναι τόσο



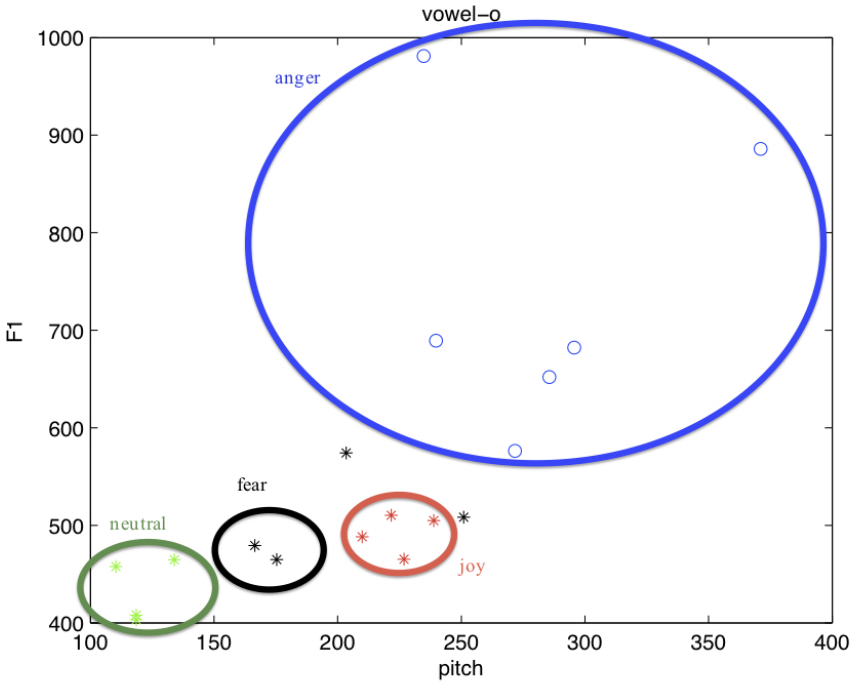
Σχήμα 4.7: Box plots των 4 πρώτων formants για τα 7 συναισθήματα.

σαφής. Παρόλα αυτά ο θυμός και το ουδέτερο διατηρούν και σε αυτήν την περίπτωση τη διακριτότητά τους.

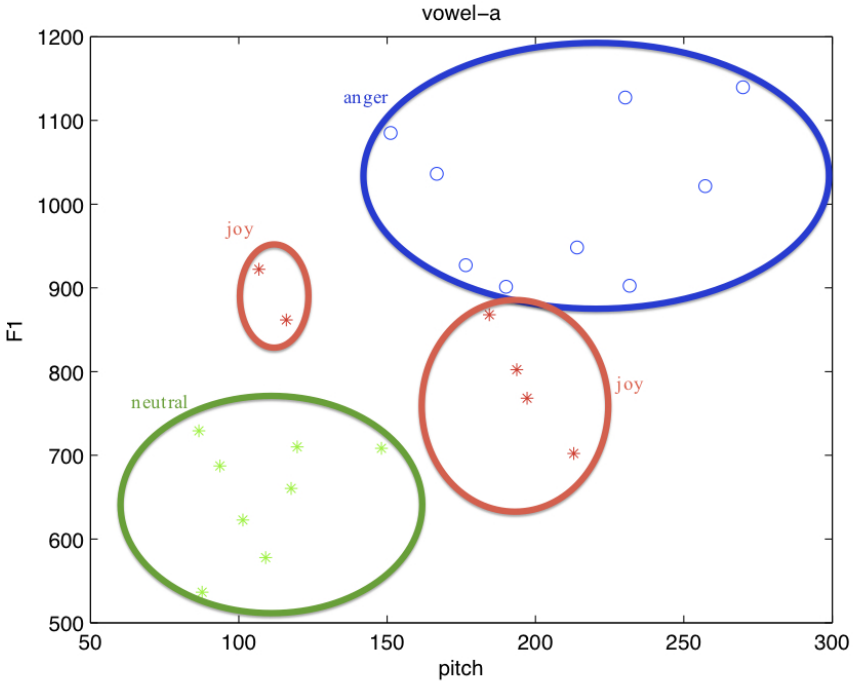
Για να ενισχύσουμε την πληροφορία που παίρνουμε από τις διαμορφώτριες συχνότητες, προσθέτουμε την τιμη του pitch. Απεικονίζουμε γραφικά την τιμή του pitch και του F1 στον οριζόντιο και τον κατακόρυφο άξονα αντίστοιχα για τα 4 φωνήεντα, όπως φαίνεται στα σχήματα 4.9 έως 4.12. Παρατηρούμε ότι στο 'a' και στο 'ö' διακρίνονται καλύτερα τα συναισθήματα απ'οτι στο 'e' και το 'ie'. Τα φωνήεντα 'a' και 'ö' χαρακτηρίζονται ως "back vowels" που σημαίνει ότι προφέρονται με τη γλώσσα να βρίσκεται στο πίσω μέρος του στόματος. Το συναίσθημα ίσως έχει μεγαλύτερη επίδραση στην κίνηση αυτή, απ'οτι αν τοποθετηθεί η γλώσσα στο μπροστινό μέρος του στόματος, οπότε και προκύπτουν τα "front vowels" ('e', 'ie'). Η παρατήρηση αυτή έρχεται σε συμφωνία με προηγούμενη έρευνα [96], όπου διαπιστώθηκε ότι η διαφορά μεταξύ των κλάσεων των 3 συναισθημάτων (θυμός, χαρά, λύπη) και του ουδέτερου ως προς τις διαμορφώτριες συχνότητες είναι πιο εμφανής στα back vowels.



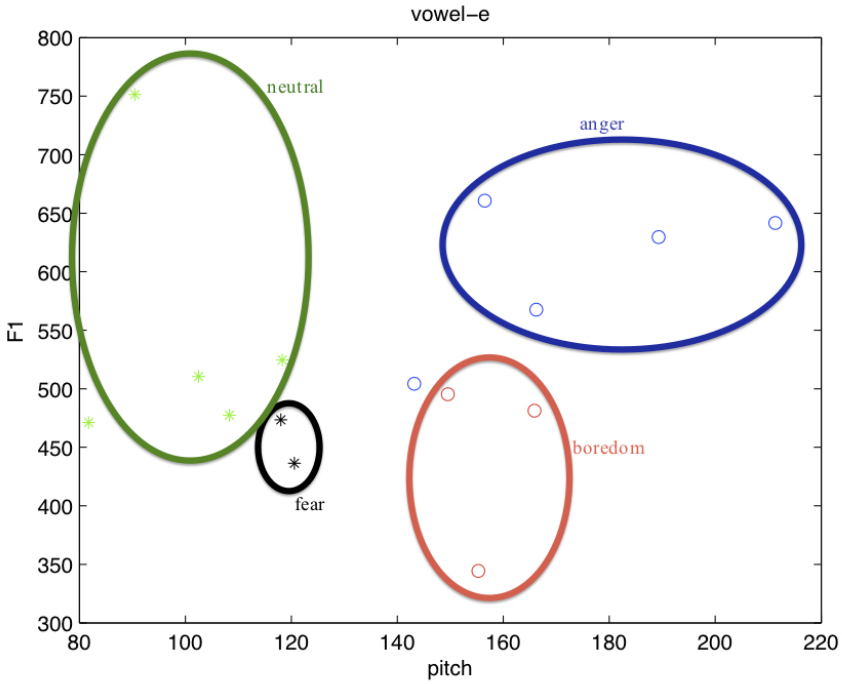
Σχήμα 4.8: Γραφική παράσταση των ζευγών a)F1 και F2 και b)F2 και F3 για τα φωνήματα 'a'(o), 'e'(x), 'ie'(+) σε 3 συναισθήματα.



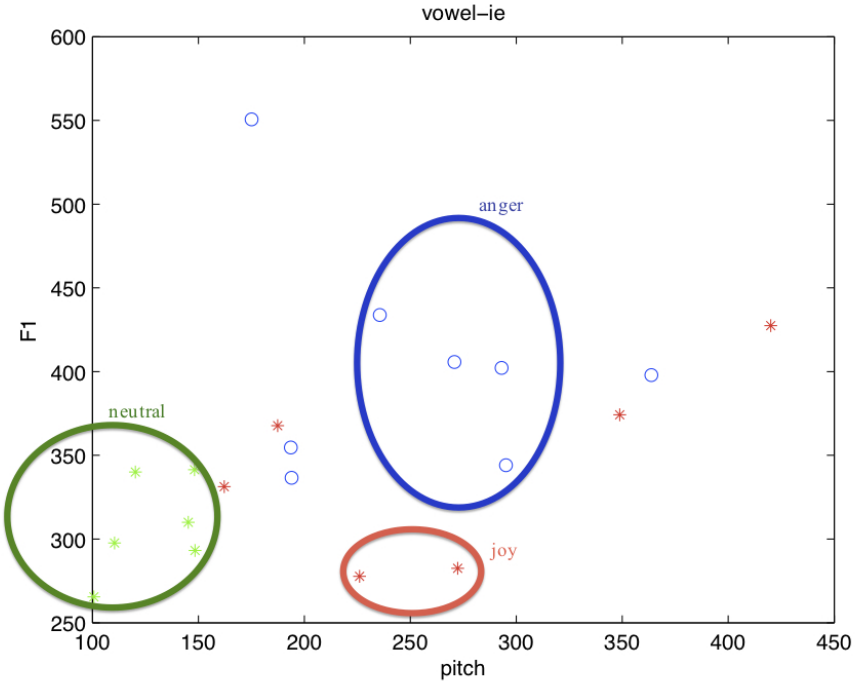
Σχήμα 4.9: Γραφική παράσταση των ζευγών pitch και F1 για το φώνημα "o" σε 4 συναισθήματα.



Σχήμα 4.10: Γραφική παράσταση των ζευγών pitch και F1 για το φώνημα "a" σε 3 συναισθήματα.



Σχήμα 4.11: Γραφική παράσταση των ζευγών pitch και F1 για το φώνημα "e" σε 4 συναισθήματα.



Σχήμα 4.12: Γραφική παράσταση των ζευγών pitch και F1 για το φώνημα "ie" σε 3 συναισθήματα.

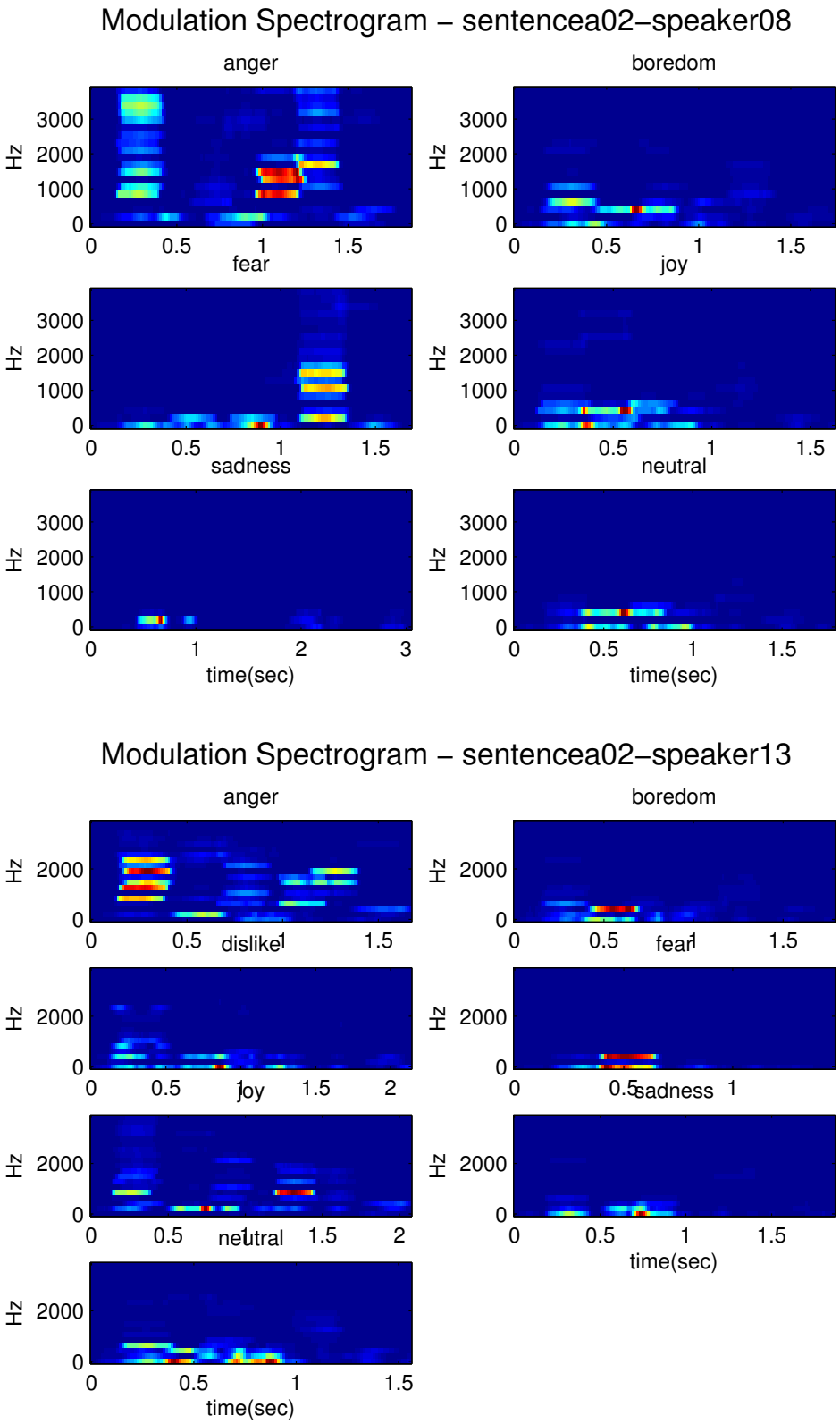
4.4 Συντελεστές Χαμηλών Συχνοτήτων Μετασχηματισμού **Fourier**

Οι συντελεστές χαμηλών συχνοτήτων του μετασχηματισμού Fourier απεικονίζονται μέσω του modulation spectrogram. Το modulation spectrogram [32] είναι μία αναπαράσταση των αργών διαμορφώσεων του σήματος φωνής στο χρόνο και τη συχνότητα. Το διάγραμμα αυτό δείχνει τις συχνότητες διαμόρφωσης του σήματος φωνής μεταξύ 0 και $8Hz$ με υψηλότερη ευαισθησία στα $4Hz$. Οι συχνότητες διαμόρφωσης υπολογίζονται στο διάστημα $0-4000Hz$ για 19 κανάλια με αυξανόμενο έυρος.

Το σήμα ομιλίας διαχωρίζεται στα 19 κανάλια μέσω ενός filterbank τραπεζοειδών φίλτρων με μηδενική επικάλυψη μεταξύ τους. Σε κάθε κανάλι το φιλτραρισμένο σήμα υφίσταται μονόδρομη ανόρθωση και βαθυπερατό φιλτράρισμα, ώστε να παραχθεί η περιβάλλουσά του. Στη συνέχεια, γίνεται υποδειγματοληψία από $16000Hz$ σε $80Hz$ και κανονικοποίηση με τη μέση τιμή. Το σήμα που προκύπτει από την παραπάνω διαδικασία χωρίζεται σε frames διάρκειας $250ms$ με βήμα $12.5ms$, ώστε να διακρίνονται τα δυναμικά χαρακτηριστικά του. Σε κάθε frame υπολογίζεται ο Fourier μετασχηματισμός και απεικονίζονται τα πλάτη αυτού που αντιστοιχούν στα $4Hz$ για κάθε frame και κάθε κανάλι.

Υποστηρίζεται ότι το modulation spectrogram είναι πιο ευσταθές από το συμβατικό φασματόγραμμα σε σχέση με το θόρυβο και τις συνθήκες αντίχησης.

Στο σχήμα 4.13 απεικονίζεται το modulation spectrogram της πρότασης a02 για 6 διαφορετικά συναισθήματα, για τα οποία παρατηρούμε αρκετές διαφορές. Ο θυμός παρουσιάζει πολλές συχνότητες διαμόρφωσης $4Hz$ σε όλες τις συχνότητες και σε ευρύ χρονικό διάστημα, σε αντίθεση με το θυμό που έχι υψηλές τιμές σε μεμονωμένη χρονική στιγμή. Η πλήξη, η χαρά και το ουδέτερο έχουν υψηλές τιμές σε μεσαίες συχνότητες, ενώ η λύπη σχεδόν σε μηδενικές συχνότητες.



Σχήμα 4.13: Modulation Spectrogram της πρότασης "Das will sie am Mittwoch abgeben." ("She will hand it in on Wednesday") για τους ομιλητές 08 και 13.

4.5 Συμπεράσματα

Στο κεφάλαιο αυτό μελετήσαμε τα πιο συνηθισμένα χαρακτηριστικά για αναγνώριση συναισθήματος.

Η χρονική διάρκεια εκφοράς, παρ΄ ότι δεν είναι το πιο διακριτικό χαρακτηριστικό, βλέπουμε ότι παίζει ρόλο στη διαμόρφωση του συναισθήματος. Αν συγκρίνουμε τη διάρκεια εκφοράς μίας συγκεκριμένης πρότασης, θα δούμε ότι διαφοροποιείται για κάθε συναίσθημα, όπως φάνηκε στο διάγραμμα του σχήματος 4.2. Η απερέσκεια και η λύπη εκφράζονται με προτάσεις μεγαλύτερης διάρκειας, ενώ ο φόβος και η χαρά με μικρότερες προτάσεις.

Το pitch και η προσωδία είναι τα πιο ευρέως χρησιμοποιούμενα χαρακτηριστικά στις έρευνες αναγνώρισης συναισθήματος. Όπως παρατηρήσαμε στο σχήμα 4.3, ο θυμός και η χαρά έχουν υψηλές τιμές pitch, ενώ η πλήξη και η λύπη έχουν χαμηλές τιμές. Η μεταβολή του pitch, που αντιστοιχεί στην προσωδία, είναι επίσης σημαντική. Μεγάλη μεταβλητότητα παρουσιάζουν ο θυμός και η χαρά, ενώ μικρή μεταβλητότητα η πλήξη και η λύπη. Μπορούμε λοιπόν να συμπεράνουμε ότι τα συναισθήματα που εξωτερικεύονται πιο έντονα, όπως ο θυμός και η χαρά, εκφράζονται με μεγαλύτερο και απότομα μεταβαλλόμενο pitch, ενώ αυτά που βιώνονται εσωτερικά έχουν χαμηλότερο και πιο ομαλά μεταβαλλόμενο pitch.

Τα formants αντικατοπτρίζουν, επίσης, αρκετή πληροφορία συναισθήματος, που δε φαίνεται στο pitch. Αυτό γίνεται εμφανες στα σχήματα 4.9 έως 4.12, όπου τα συναισθήματα δε θα μπορούσαν να διαχωριστούν μόνο μέσω του pitch. Τα formants και ειδικότερα το 1ο και το 2ο εμπεριέχουν πληροφορία για τη μορφή της στοματικής κοιλότητας, που επηρεάζεται πολύ από το συναίσθημα.

Τέλος, οι χαμηλόσυχνοι συντελεστές του μετασχηματισμού Fourier παρουσιάζουν αρκετά μεγάλη διακριτότητα μεταξύ των συναισθημάτων και μπορεί να αποτελέσουν ένα απλό και ταυτόχρονα αποδοτικό χαρακτηριστικό.

Κεφάλαιο 5

Χαρακτηριστικά Διαμόρφωσης AM-FM

Έχουν προταθεί πολλοί τρόποι για τη μοντελοποίηση της ομιλίας. Μία από αυτές είναι η μοντελοποίηση του αέρα στο φωνητικό σωλήνα, που συντελεί στη διαμόρφωση των φωνημάτων στην ομιλία [79]. Οι μελέτες που θεωρούν το σήμα φωνής ως ένα μονοδιάστατο κύμα και εστιάζουν στη γραμμική διάδοσή του στο φωνητικό σωλήνα έχουν περιορισμένη απόδοση. Αντίθετα, υπάρχουν ενδείξεις ότι το σήμα φωνής μπορεί να θεωρηθεί ως ένα τρισδιάστατο μη γραμμικό δυναμικό φαινόμενο [25].

Ο Teager τελεστής είναι ένας μη γραμμικός ενεργειακός τελεστής που ανακλά τη διάδοση του αέρα στο φωνητικό σωλήνα και δίνεται από τη σχέση:

$$\Psi[x(t)] = (\frac{d}{dt}x(t))^2 - x(t)(\frac{d^2}{dt^2}x(t)) \ = [\dot{x}(t)]^2 - x(t)\ddot{x}(t)$$

Ο αλγόριθμος διαχωρισμού ενέργειας (ESA) χρησιμοποιεί τον Teager τελεστή και εντοπίζει τις διαμορφώσεις στο σήμα φωνής υπολογίζοντας το στιγμιαίο πλάτος και τη στιγμιαία συχνότητα με βάση τις σχέσεις:

$$|\alpha(t)| = \frac{\Psi[x(t)]}{\sqrt{\Psi[\dot{x}(t)]}}$$

$$\omega(t) = \sqrt{\frac{\Psi[\dot{x}(t)]}{\Psi[x(t)]}}$$

Έχοντας παραθυροποιήσει το σήμα φωνής, σε κάθε παράθυρο εφαρμόζουμε 4,13, 20 και 30 Gabor φίλτρα, οπότε προκύπτουν και τα αντίστοιχα κανάλια συχνότητας για κάθε παράθυρο. Τα φίλτρα είναι κατανεμημένα στην κλίμακα Mel και σε εύρος συχνοτήτων 0-4000Hz και 0-8000Hz. Σε κάθε κανάλι υπολογίζουμε το μέσο όρο και την τυπική απόκλιση του στιγμιαίου πλάτους $\alpha(t)$ και της στιγμιαίας συχνότητας $\omega(t)$:

$$\alpha_{mean} = \frac{\sum_{t=1}^{N} |\alpha(t)|}{N}$$

$$\alpha_{std} = \sqrt{\frac{\sum_{t=1}^{N}(\alpha(t) - \alpha_{mean})^2}{N-1}}$$

$$\omega_{mean} = \frac{\sum_{t=1}^{N}|\omega(t)|}{N}$$

$$\omega_{std} = \sqrt{\frac{\sum_{t=1}^{N}(\omega(t) - \omega_{mean})^2}{N-1}}$$

όπου N: το μέγεθος κάθε καναλιού.

Επίσης, για να αναδειχθούν καλύτερα οι μεταβολές της στιγμιαίας συχνότητας, υπολογίζουμε τα μεγέθη F, B και FMP Το F είναι ο σταθμισμένος μέσος όρος της στιγμιαίας συχνότητας με βάρη το στιγμαίο πλάτος.

$$F = \frac{\sum_{t=1}^{N}\omega(t)\alpha^2(t)}{\sum_{t=1}^{N}\alpha^2(t)}$$

Το B είναι μία σταθμισμένη εκδοχή της τυπικής απόκλισης της στιγμιαίας συχνότητας.

$$B = \frac{\sum_{t=1}^{N}[\dot{\alpha}^2(t) + (\omega(t) - F)^2\alpha^2(t)]}{\sum_{t=1}^{N}\alpha^2(t)}$$

Το FMP είναι το ποσοστό διαμόρφωσης στη συχνότητα (Frequency Modulation Percentage) και υπολογίζεται ως ο λόγος των μεγεθών B δια F [20].

$$FMP = \frac{B}{F}$$

Για να απαλείψουμε τις απότομες μεταβολές των ESA χαρακτηριστικών στο χρόνο, βρίσκουμε την περιβάλλουσα του στιγμιαίου πλάτους και στη συνέχεια υπολογίζουμε το εμβαδόν της αυτοσυσχέτισής της. Η κανονικοποιημένη τιμή του εμβαδού αυτού είναι το TEO-Auto-Env χαρακτηριστικό.

$$TEO - Auto - Env = \frac{\sum_{t=1}^{2N-1}(\alpha(t) * \alpha(t))_{env}}{2N-1}$$

όπου $(\alpha(t) * \alpha(t))_{env}$ είναι η περιβάλλουσα της αυτοσυσχέτισης του στιγμιαίου πλάτους.
Αντί να υπολογίσουμε το εμβαδό της καμπύλης της περιβάλλουσας, μπορούμε να εφαρμόσουμε πολυωνυμική παρεμβολή της καμπύλης αυτής [33]. Με τον τρόπο αυτό βρίσκουμε τους συντελεστές ενός πολυωνύμου 3ου βαθμού $p(X) = C_0 + C_1X + C_2X^2 + C_3X^3$. Η περιβάλλουσα της αυτοσυσχέτισης του στιγμιαίου πλάτους θεωρείται γραμμική, ενώ οι όροι 2ης και 3ης τάξης προστίθενται για μεγαλύτερη ακρίβεια.
Τέλος, λαμβάνοντας υπόψη ότι ο τελεστής TEO έχει την ίδια περιοδικότητα για όλα τα χαρακτηριστικά που προκύπτουν με βάση αυτόν [98], υπολογίζουμε το pitch της στιγμιαίας συχνότητας $\omega(t)$ και αυτό ονομάζουμε TEO-Pitch.

$$TEO - Pitch = pitch(\omega(t))$$

Τα AM-FM χαρακτηριστικά διαμόρφωσης έχουν ίδια περιοδικότητα με το pitch. Το πλεονέκτημα του TEO-Pitch σε σχέση με το pitch είναι ότι εντοπίζει καλύτερα την περιοδικότητα του αρχικού σήματος λόγω του τετραγωνισμού στο πλάτος και τη συχνότητα που εφαρμόζεται στον Teager τελεστή.
Για να εξετάσουμε τη χρονική εξέλιξη των χαρακτηριστικών, μπορούμε να υπολογίσουμε την παράγωγο του στιγμιαίου πλάτους με δυο τρόπους. Ο πρώτος τρόπος είναι ο πιο τετριμμένος, αφού υπολογίζει τη διαφορά μεταξύ διαδοχικών δειγμάτων του χαρακτηριστικού αυτού, δηλαδή:

$$Ampl - Der1(t) = \alpha(t+1) - \alpha(t)$$

Ο δεύτερος τρόπος χρησιμοποιεί τον ορισμό της παραγώγου μέσω της γκαουσιανής συνάρτησης και δίνεται από τη σχέση:

$$Ampl - Der2(t) = \lim_{\sigma \to 0} \frac{d}{dt}\alpha(t) * G_\sigma(t) = \lim_{\sigma \to 0} \alpha(t) * \frac{d}{dt}G_\sigma(t)$$

Σε κάθε παράθυρο υπολογίζονται ο μέσος όρος και η τυπική απόκλιση των παραγώγων του στιγμιαίου πλάτους, όπως ακριβώς και για το στιγμιαίο πλάτος και τα συμβολίζουμε με $Ampl - Der1_{mean}$, $Ampl - Der1_{std}$, $Ampl - Der2_{mean}$ και $Ampl - Der2_{std}$.

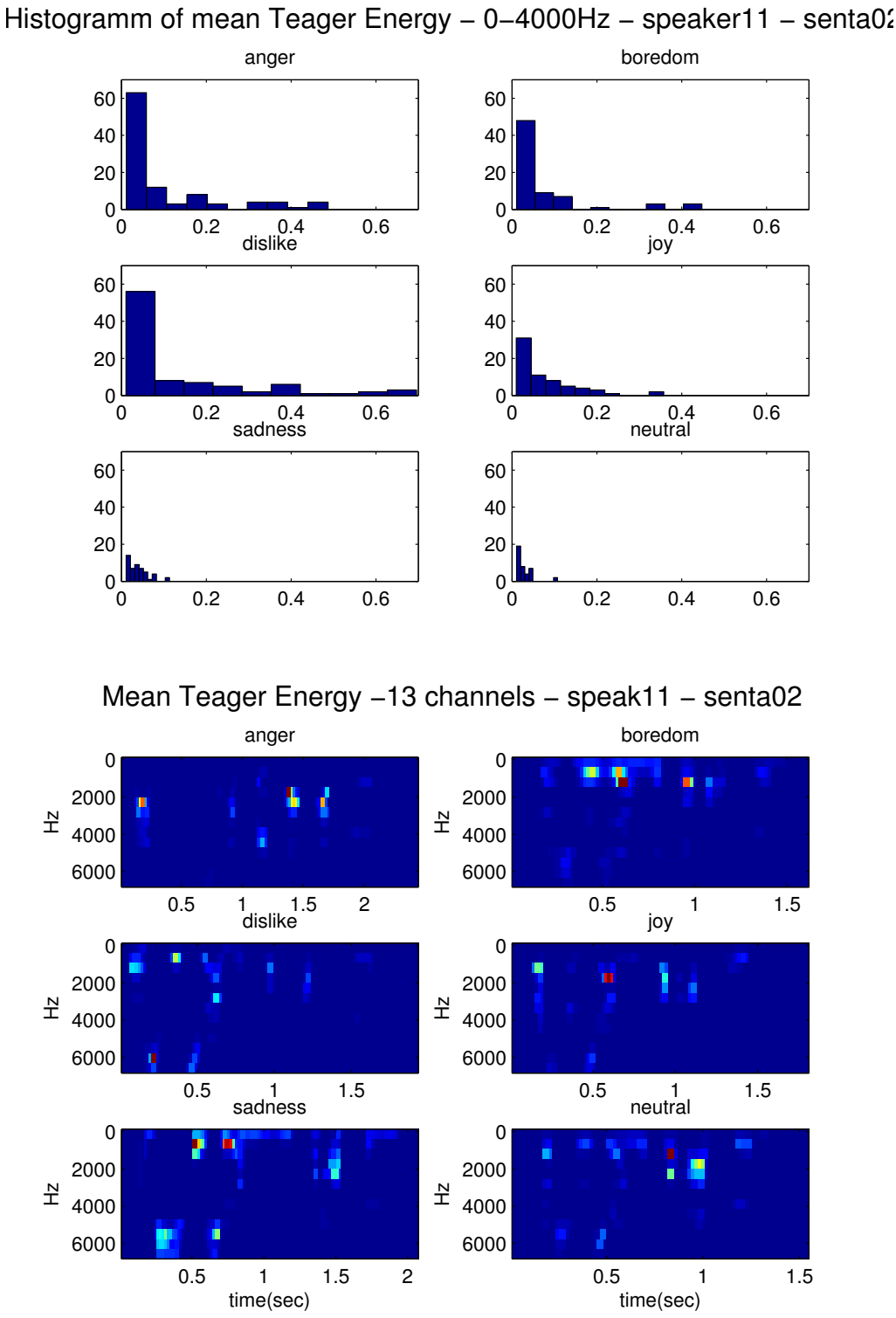
5.1 Τελεστής Teager Ενέργειας

Η ενέργεια του σήματος φωνής χρησιμοποιείται εκτενώς στην αναγνώριση συναισθήματος. Αυτό γιατί παραπέμπει στην ένταση της ομιλίας που είναι ενδεικτική του συναισθήματος. Υπολογίζουμε μία εκδοχή ενέργειας του σήματος φωνής, που βασίζεται στον τελεστή Teager. Σχεδιάζουμε τα φασματογράμματα και τα ιστογράμματα ίδιων προτάσεων που εκφράζονται από τον ίδιο ομιλητή για τα 4 βασικά συναισθήματα (θυμό, χαρά, λύπη, φόβο) και το ουδέτερο και φαίνονται στο σχήμα 5.1. Επειδή η Teager ενέργεια έχει σε πολλά σημεία μηδενικές τιμές, στα ιστογράμματα παραλείπουμε τις τιμές αυτές, ώστε να έχουμε μία καλύτερη εποπτική εικόνα.

Σε γενικές γραμμές, παρατηρούμε ότι η απαρέσκεια και ο θυμός έχουν τις μεγαλύτερες τιμές της Teager ενέργειας, ενώ σχετικά μεγάλες τιμές έχουν και η πλήξη και η χαρά. Επίσης, η λύπη και το ουδέτερο συναίσθημα εμφανίζουν πολύ μικρές τιμές μέσης ενέργειας. Από το φασματόγραμμα βλέπουμε ότι παρά τις υψηλές τιμές σε μεμονωμένα σημεία, ο θυμός έχει περισσότερες μηδενικές τιμές από τα υπόλοιπα συναισθήματα σε όλη την έκταση χρόνου και συχνότητας. Αντίθετα, στη χαρά και στο φόβο παρατηρούνται διάσπαρτες μη μηδενικές τιμές. Στη λύπη οι μη μηδενικές τιμές βρίσκονται σε μεσαίες και υψηλές ζώνες συχνοτήτων και συνήθως εμφανίζουν πιο μακρά και συνεχόμενη χρονική διάρκεια, που αντιστοιχεί σε μεγάλα οριζόντια φωτεινά τμήματα στα φασματογράμματα.

5.2 Στιγμαίο Πλάτος

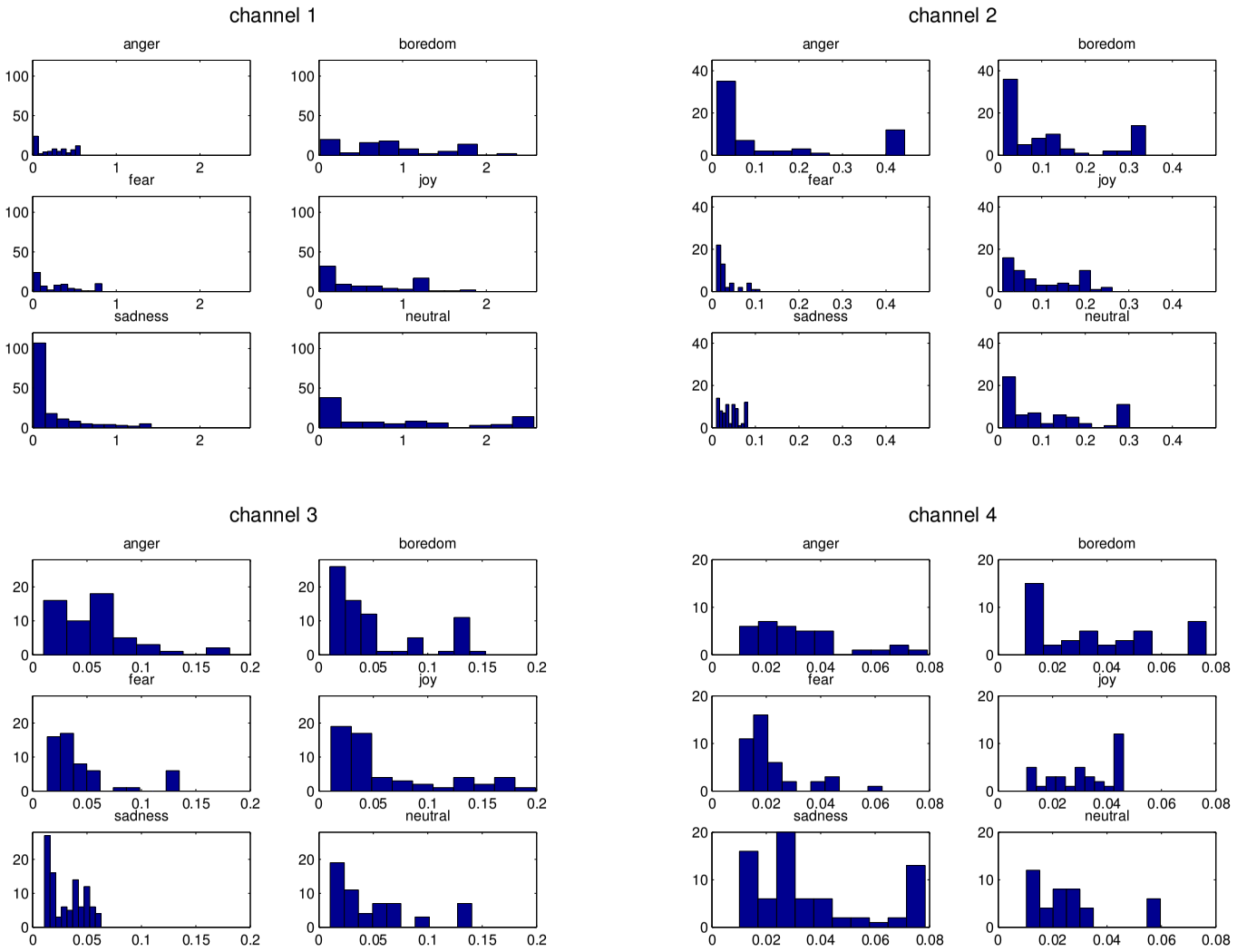
Εξετάζουμε την κατανομή του μέσου στιγμιαίου πλάτους με τη βοήθεια ιστογραμμάτων. Υπολογίζουμε το στιγμιαίο πλάτος κάθε πρότασης σε 4 κανάλια και απεικονίζουμε τα ιστογράμματα των τιμών του. Στα ιστογράμματα αυτά απομακρύνουμε τις τιμές μικρότερες του 0.1, που οφείλονται σε διαίρεση με πολύ υψηλές τιμές από τον αλγόριθμο ESA, ώστε να



Σχήμα 5.1: Ιστόγραμμα και φασματόγραμμα των τιμών μέσης Teager ενέργειας της πρότασης "Das will sie am Mittwoch abgeben." ("She will hand it in on Wednesday.") στα 4 βασικά συναισθήματα (θυμός, χαρά, λύπη, φόβος) και στο ουδέτερο.

υπάρχει μία πιο καθαρή απεικόνιση. Επίσης, το στιγμιαίο πλάτος έχει υποστεί alpha-mean trimming, με σκοπό να εξαλειφθούν οι outliers. Στο σχήμα 5.2 βλέπουμε το ιστόγραμμα του στιγμιαίου πλάτους για τον ομιλητή 08 και την πρόταση a02. Παρατηρούμε ότι η λύπη έχει πιο πολλές τιμές κοντά στο 0 σε σχέση με τα υπόλοιπα συναισθήματα στο 1ο κανάλι. Αυτό δικαιολογείται από το γεγονός ότι η ένταση της φωνής είναι μειωμένη στη λύπη. Επίσης, αξιοσημείωτο είναι ότι ο θυμός έχει μικρές τιμές πλάτους στις χαμηλότερες συχνότητες, σε αντίθεση με την πλήξη και το ουδέτερο που παρουσιάζουν πιο αυξημένες τιμές στο πρώτο κανάλι. Στα υπόλοιπα κανάλια και ειδικότερα στο 2ο και το 3ο, που αντιπροσωπεύουν μεσαίες προς υψηλές συχνότητες, ο θυμός έχει πιο αυξημένο στιγμιαίο πλάτος από τα υπόλοιπα συναισθήματα. Η λύπη παρουσιάζει στα 3 πρώτα κανάλια χαμηλές τιμές, αλλά στο 4ο κανάλι φαίνεται να έχει τις περισσότερες υψηλές τιμές από όλα τα συναισθήματα.

Στα σχήματα 5.3 και 5.4 φαίνονται οι γραφικές παραστάσεις της χρονικής εξέλιξης των ροπών 1ης έως 4ης τάξης για το στιγμαίο πλάτος στα $0-1000Hz$. Όπως αναμένεται και από τα ιστογράμματα, ο θυμός, η χαρά και η απαρέσκεια έχουν χαμηλές τιμές στιγμαίου πλάτους, ενώ η πλήξη και ο φόβος υψηλότερες τιμές. Η κατανομή του στιγμαίου πλάτους για όλα τα συναισθήματα είναι ασύμμετρη προς τα αριστερά. Ωστόσο, παρατηρούμε ότι η κύρτωση είναι το μέγεθος που παρουσιάζει περισσότερες διαφορές. Μεγάλες τιμές κύρτωσης φαίνονται στο θυμό και την απαρέσκεια, πράγμα που υποδηλώνει ότι τα δύο αυτά συναισθήματα έχουν πιο πολλές ασυνήθιστες αποκλίσεις τιμών. Αντίθετα τα υπόλοιπα συναισθήματα εμφανίζουν κύρτωση κοντά στο 3, τιμή που αντιστοιχεί στη γκαουσιανή κατανομή.

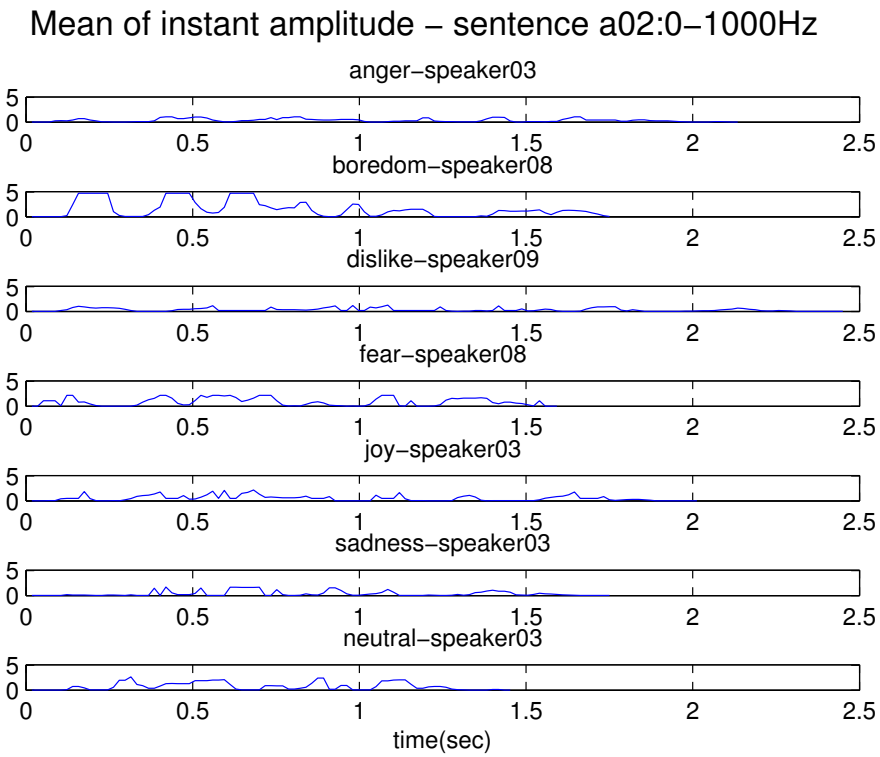


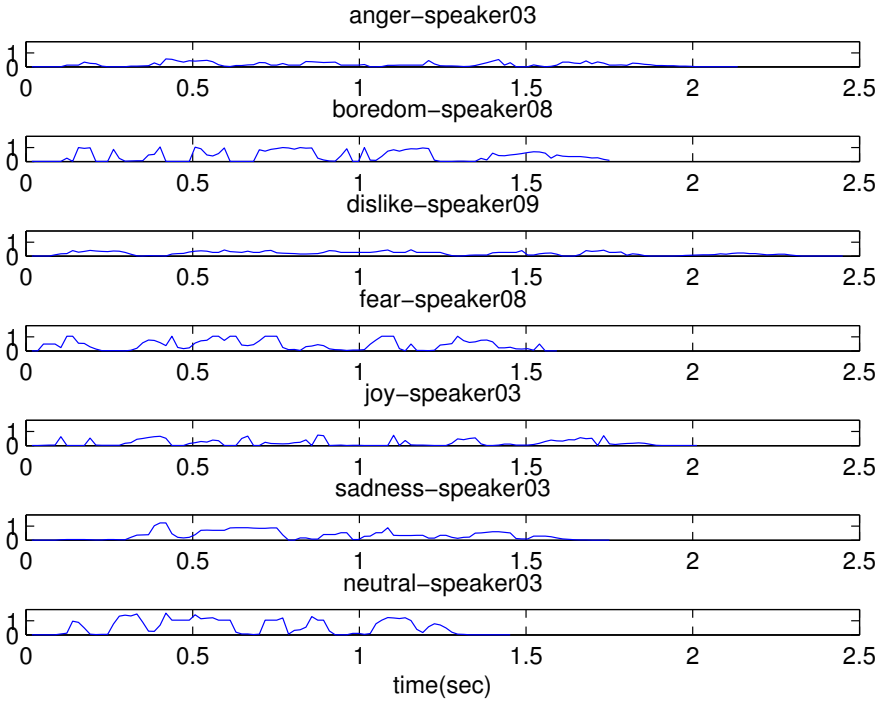
Σχήμα 5.2: Ιστογράμματα του στιγμιαίου πλάτους της πρότασης a02: "Das will sie am Mittwoch abgeben." ("She will hand it in on Wednesday.") για 6 συναισθήματα σε 4 κανάλια.

Στη συνέχεια, υπολογίζουμε την παράγωγο του στιγμιαίου πλάτους, ώστε να διαπιστώσουμε τις μεταβολές του μεγέθους αυτού για κάθε συναίσθημα. Το στιγμιαίο πλάτος υπολογίζεται πάλι στο φωνήεν 'a' για 6 συναισθήματα. Η παράγωγος του πλάτους βρίσκεται ως η διαφορά των τιμών σε διαδοχικές χρονικές στιγμές. Στα σχήματα 5.5 έως και 5.6 απεικονίζεται το ζωνοπερατό φιλτράρισμα του αρχικού σήματος του φωνήεντος, το στιγμαίο πλάτος και η παράγωγος αυτού. Στο σχήμα της παραγώγου και για κάθε frame φαίνεται η μέση τιμή και η διαφορά της τυπικής απόκλισης από τη μέση τιμή (με κόκκινο και πράσινο χρώμα αντίστοιχα). Σημειώνουμε ότι έχει εφαρμοστεί alpha-trimmed mean filtering για την εξάλειψη των outliers στο στιγμιαίο πλάτος και την παράγωγό του.

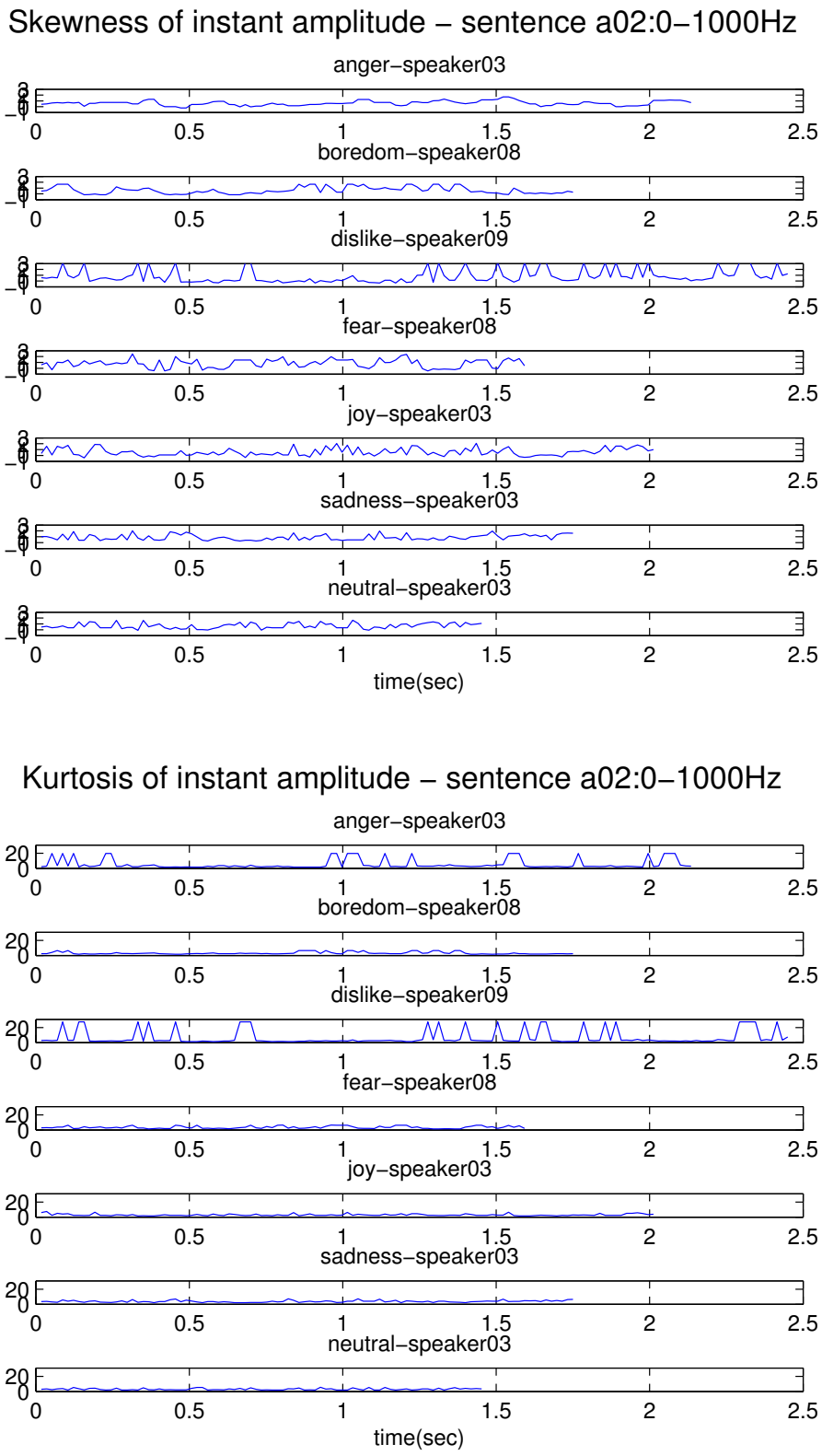
Ανάμεσα στα συναισθήματα παρατηρούνται μεγάλες διαφορές, ειδικά στην συχνότητα μεταβολής του στιγμιαίου πλάτους τους. Στο θυμό και στη λύπη το στιγμιαίο πλάτος μεταβάλλεται με πολύ γρήγορο ρυθμό, ενώ στα υπόλοιπα συναισθήματα βλέπουμε πιο ομαλές μεταβολές. Η παρατήρηση αυτή αντικατοπρίζεται και στην παράγωγο του στιγμαίου πλάτους. Παρά τις γρήγορες μεταβολές, ο θυμός παρουσιάζει μικρές τιμές στιγμιαίου πλάτους, σε αντίθεση με τα άλλα συναισθήματα που έχουν μεγαλύτερες τιμές. Επίσης, η πλήξη έχει σχεδόν σταθερό μέσο όρο στιγμαίου πλάτους ανάμεσα στα frames. Αυτό δε συμβαίνει για το φόβο και τη χαρά, όπου το στιγμαίο πλάτος παρουσιάζει φθίνουσα πορεία με την πάροδο του χρόνου.

Στο σχήμα 5.8 φαίνονται τα box plots για το μέσο όρο και την τυπική απόκλιση της παραγώγου του στιγμαίου πλάτους. Το στιγμαίο πλάτος έχει υπολογιστεί σε 4 κανάλια για 3 λέξεις (Eisschrank, Lappen, Liegt), που έχουν χωριστεί σε frames διάρκειας 10msec με 50% επικάλυψη. Παρατηρούμε ότι αυτό που διαφέρει μεταξύ των συναισθημάτων είναι οι τιμές της τυπικής απόκλισης, ενώ οι τιμές του μέσου όρου δεν έχουν μεγάλες διαφορές. Στο πρώτο κανάλι με τις χαμηλότερες συχνότητες η λύπη παρουσιάζει μεγαλύτερη μεταβολή στην παράγωγο του στιγμαίου πλάτους από τα υπόλοιπα συναισθήματα. Το αντίστοιχο συμβαίνει για τη χαρά στο 4ο κανάλι.

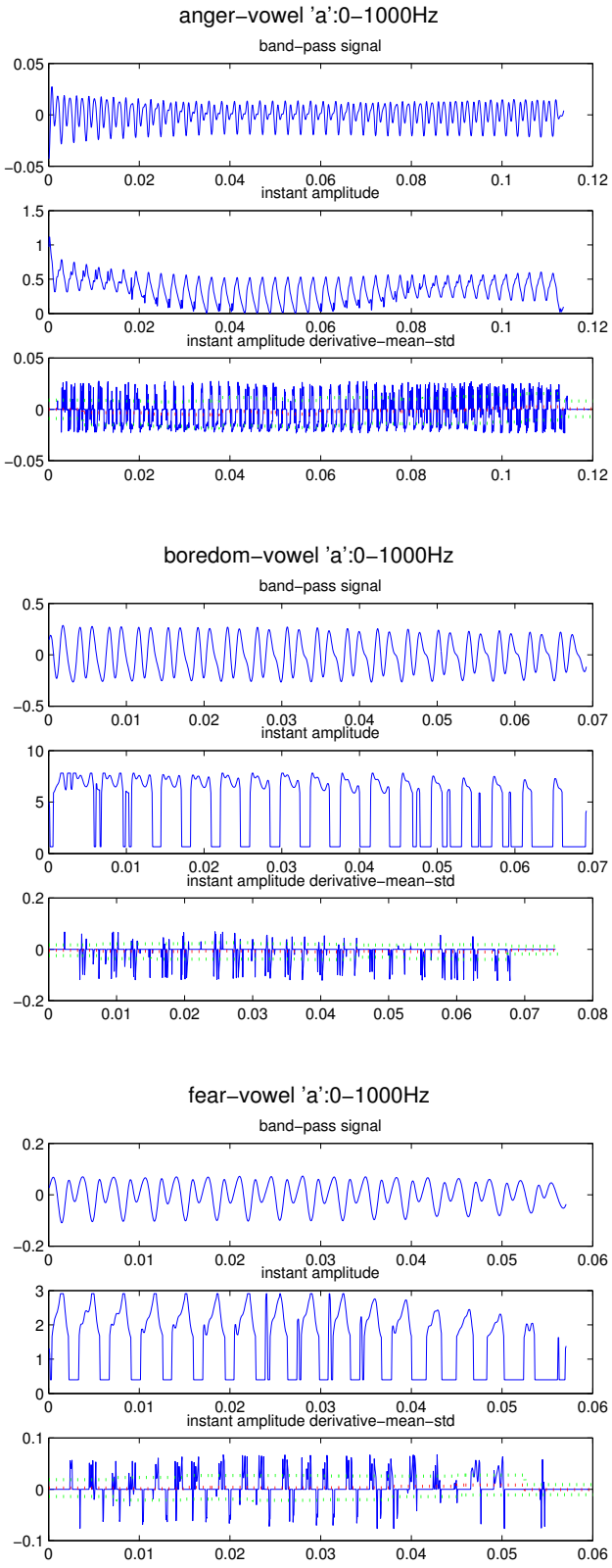




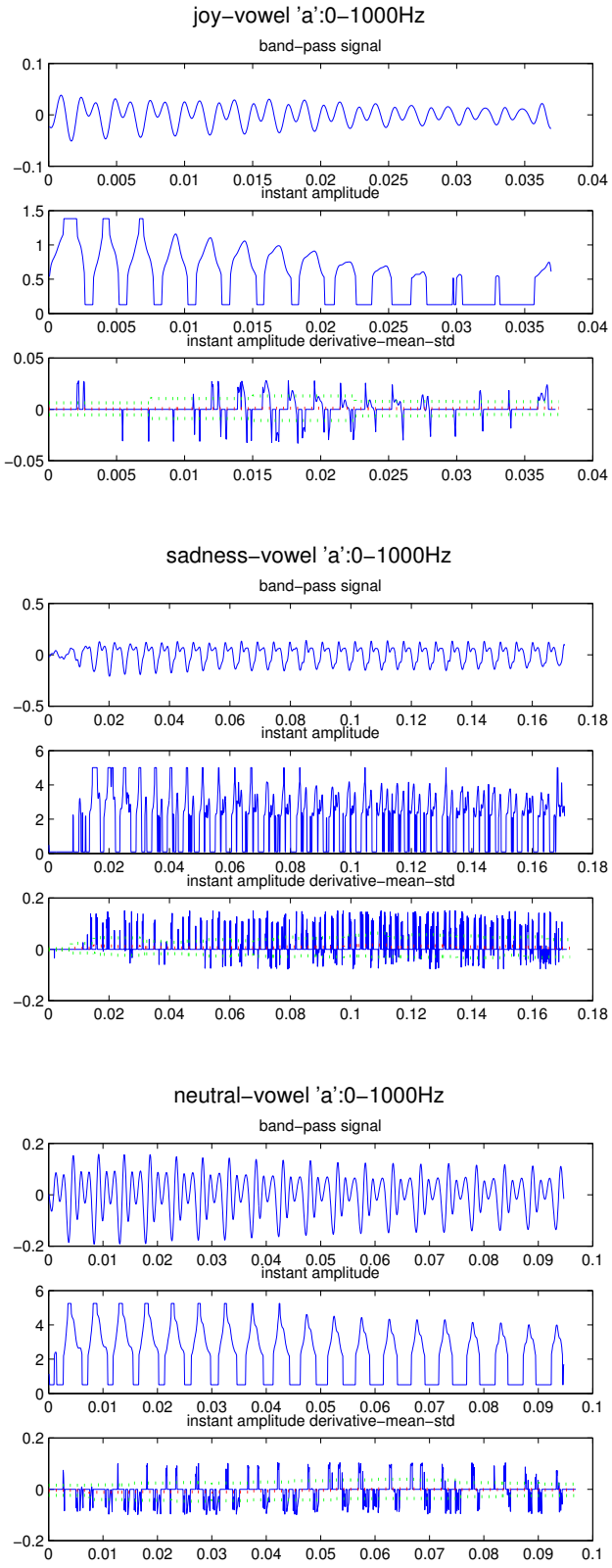
Σχήμα 5.3: Γραφικές παραστάσεις της ροπής 1ης και 2ης τάξης του στιγμιαίου πλάτους της πρότασης a02: "Das will sie am Mittwoch abgeben." ("She will hand it in on Wednesday.") για 6 συναισθήματα στο 1ο κανάλι.



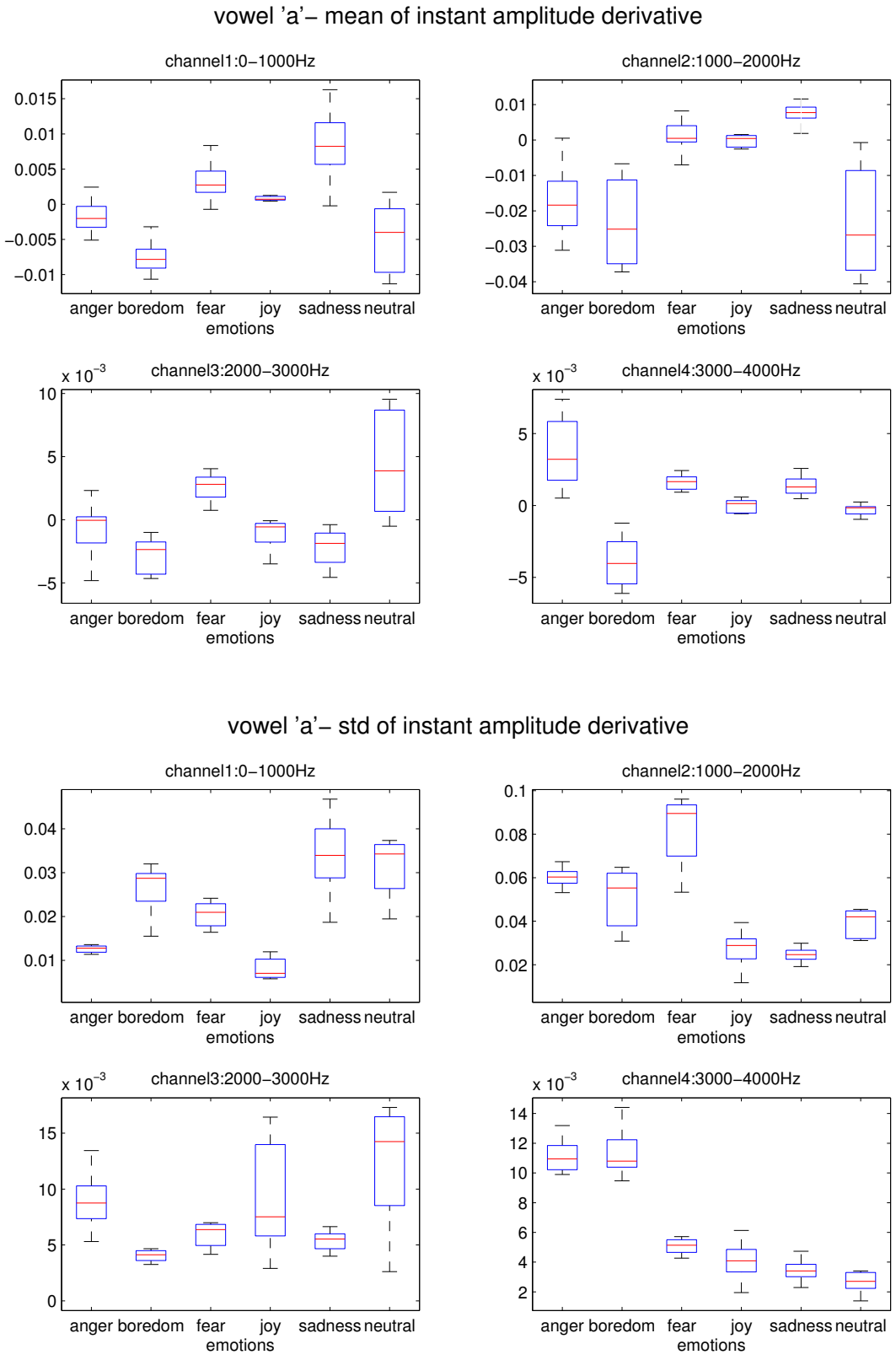
Σχήμα 5.4: Γραφικές παραστάσεις της ροπής 3ης και 4ης τάξης του στιγμιαίου πλάτους της πρότασης a02: "Das will sie am Mittwoch abgeben." ("She will hand it in on Wednesday.") για 6 συναισθήματα στο 1ο κανάλι.



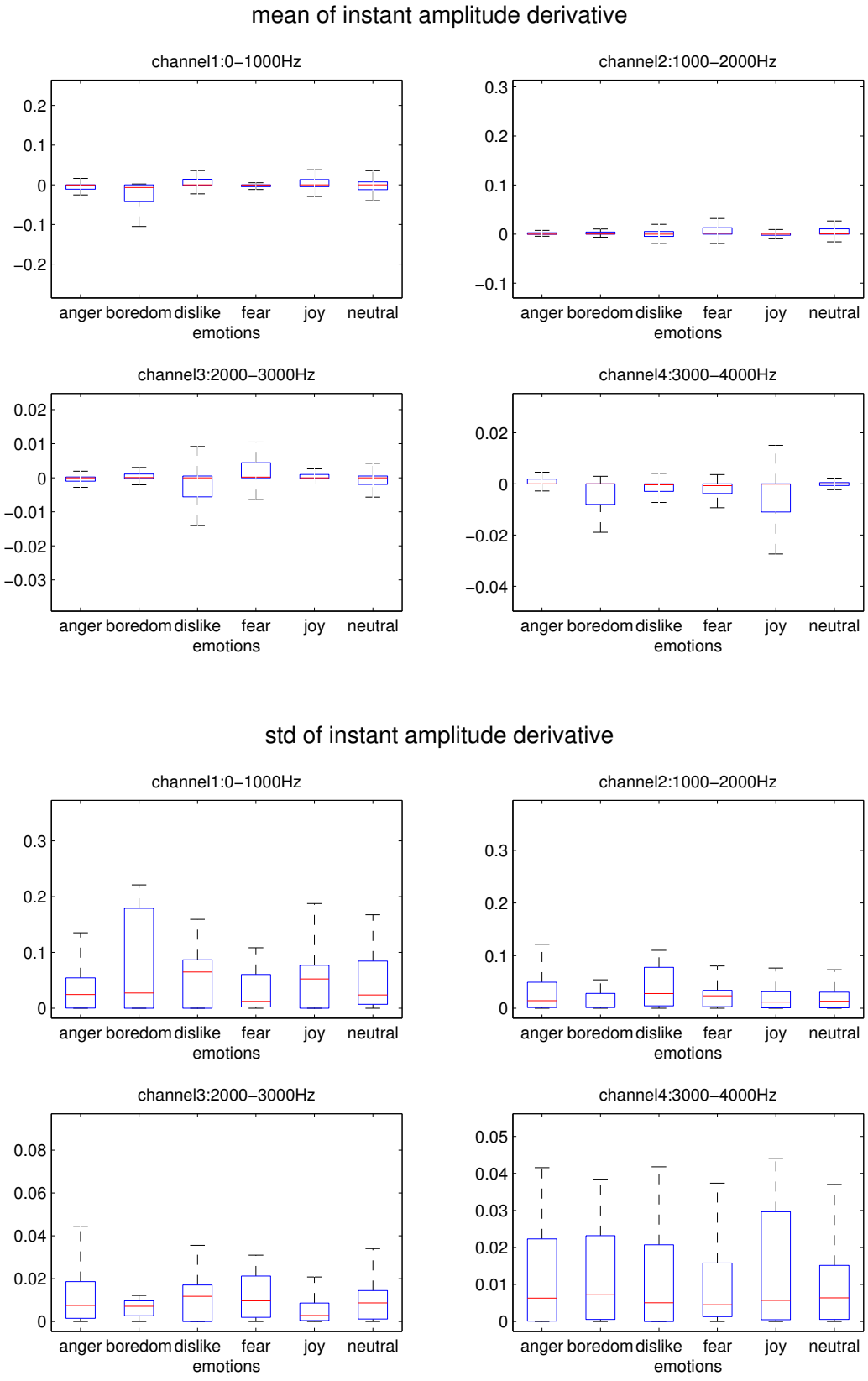
Σχήμα 5.5: Γραφική απεικόνιση του ζωνοπερατού σήματος του φωνήεντος 'a', , του στιγμιαίου πλάτους και της παραγώγου του στο θυμό, την πλήξη και το φόβο. Η κόκκινη διακεκομένη γραμμή απεικονίζει το μέσο όρο σε κάθε frame και οι δύο πράσινες διακεκομένες δείχνουν την απόσταση της τυπικής απόκλισης από το μέσο όρο.



Σχήμα 5.6: Γραφική απεικόνιση του ζωνοπερατού σήματος του φωνήεντος 'a', , του στιγμιαίου πλάτους και της παραγώγου του στη χαρά, τη λύπη και το ουδέτερο. Η κόκκινη διακεκομένη γραμμή απεικονίζει το μέσο όρο σε κάθε frame και οι δύο πράσινες διακεκομένες δείχνουν την απόσταση της τυπικής απόκλισης από το μέσο όρο.



Σχήμα 5.7: Box plots του μέσου όρου και της τυπικής απόκλισης της παραγώγου του στιγμιαίου πλάτους του φωνήεντος 'a' για 6 συναισθήματα σε 4 κανάλια.



Σχήμα 5.8: Box plots του μέσου όρου και της τυπικής απόκλισης της παραγώγου του στιγμιαίου πλάτους τριών λέξεων για 6 συναισθήματα σε 4 κανάλια.

5.3 Στιγμιαία Συχνότητα

Η στιγμιαία συχνότητα είναι ένα πολύ βασικό χαρακτηριστικό στην αναγνώριση συναισθήματος, γιατί εμπεριέχει πληροφορία για το πόσο αργός ή γρήγορος είναι ο λόγος, καθώς και για το εύρος συχνοτήτων του σήματος φωνής. Επίσης, αντικατοπτρίζει τις διαμορφώτριες συχνότητες (formants), που όπως είδαμε είναι σημαντικές στο αντικείμενο αυτό. Υπολογίζουμε μία σταθμισμένη εκδοχή του μέσου όρου και της τυπικής απόκλισης της στιγμιαίας συχνότητας με βάρη τα στιγμαία πλάτη. Τα μεγέθη αυτά τα ονομάζουμε F και B αντίστοιχα και θα τα εξετάσουμε παρακάτω.

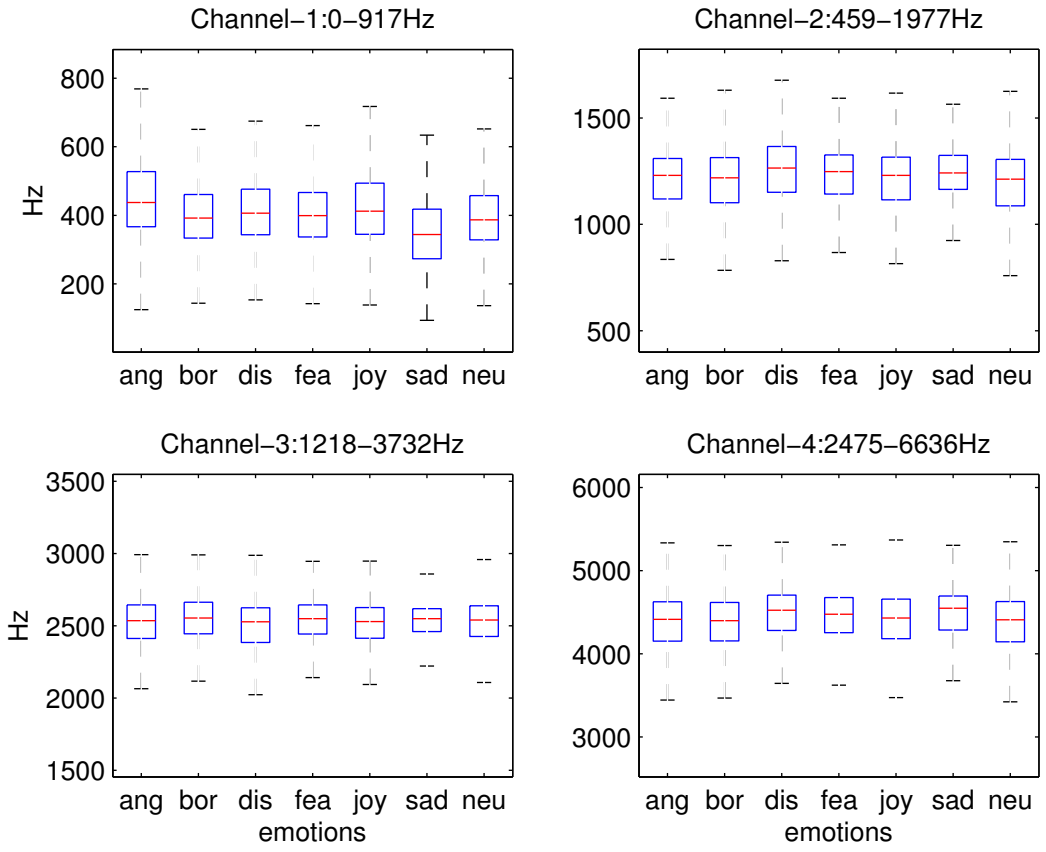
5.3.1 Σταθμισμένος Μέσος Όρος Στιγμιαίας Συχνότητας (F)

Η συχνότητα ομιλίας αποτελεί βασικό συστατικό στην αναγνώριση συναισθήματος, γιατί συμβάλει στον καθορισμό της παραγλωσσικής πληροφορίας του σήματος φωνής. Μέσω του αλγορίθμου ESA μπορεί να υπολογιστεί η στιγμιαία συχνότητα του σήματος. Σε κάθε παράθυρο μπορεί να υπολογιστεί ο απλός μέσος όρος, ο οποίος είναι υπολογιστικά γρήγορος και απλός στην έννοια. Παρόλα αυτά προτιμούμε το σταθμισμένο μέσο όρο F με βάρη τις τιμές του στιγμιαίου πλάτους, γιατί παρέχει καλύτερη εκτίμηση των διαμορφωτριών συχνοτήτων (formants) και είναι πιο εύρωστος στις χαμηλές ενέργειες καθώς και στο θόρυβο [65].

Στο σχήμα 5.9 σχεδιάζουμε τα box plots του χαρακτηριστικού F για 4 κανάλια κατανεμημένα με την κλίμακα mel στα 0-4000Hz. Οι πιο έντονες διαφορές μεταξύ των συναισθημάτων βρίσκονται στο πρώτο κανάλι. Παρατηρούμε ότι στις συχνότητες από 0 έως 917Hz το F έχει υψηλότερες τιμές και μεγαλύτερο εύρος τιμών για το θυμό και τη χαρά. Η πλήξη, η απαρέσκεια και ο φόβος έχουν χαμηλότερες τιμές με μικρότερο εύρος, ενώ η λύπη έχει τις πιο μικρές τιμές. Σχεδόν τα αντίθετα ισχύουν στις υψηλές συχνότητες από 2500Hz έως και 6000Hz. Η απαρέσκεια, ο φόβος και η λύπη παρουσιάζουν μεγαλύτερες τιμές, ενώ ο θυμός και η χαρά μικρότερες.

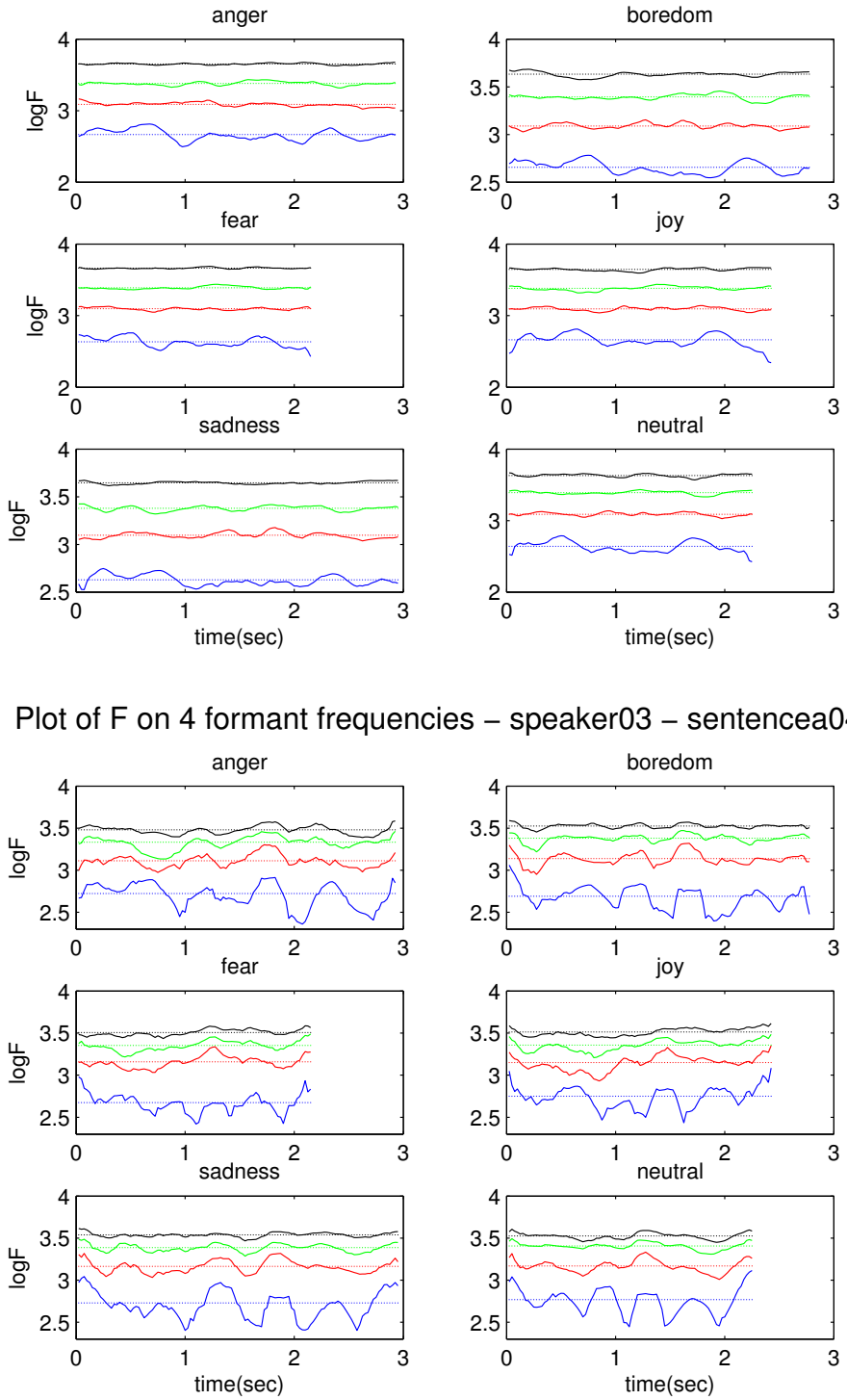
Οι παρατηρήσεις αυτές είναι εμφανείς και στο σχήμα 5.10, όπου απεικονίζεται η χρονική εξέλιξη του μεγέθους F σε 4 κανάλια για μία πρόταση και έναν ομιλητή. Το πρώτο διάγραμμα του σχήματος παριστάνει το F γύρω από προκαθορισμένες συχνότητες σύμφωνα με την κλίμακα mel, ενώ το δεύτερο διάγραμμα απεικονίζει το F γύρω από τα 4 πρώτα formants, που υπολογίζονται σε κάθε frame. Παρατηρείται ότι ο θυμός, η χαρά, ο φόβος και το ουδέτερο έχουν υψηλότερες τιμές F από τα υπόλοιπα συναισθήματα. Επίσης, στο πρώτο σχήμα βλέπουμε ότι οι υψηλές τιμές του θυμού διατηρούνται σταθερά καθόλη τη χρονική εξέλιξη του συναισθήματος αυτού. Ο σταθμισμένος μέσος όρος στιγμιαίας συχνότητας απεικονίζεται με μεγαλύτερη ευκρίνεια γύρω από τις περιοχές των 4 πρώτων formants, όπως φαίνεται και στο δεύτερο σχήμα. Η λύπη και ο θυμός έχουν μεγαλύτερες μεταβολές στις τιμές του F γύρω από το 1ο formant.

Παρόλα αυτά το χαρακτηριστικό F δε διατηρεί σε όλες τις περιπτώσεις μία συγκεκριμένη μορφή για κάθε συναίσθημα. Αναφέρουμε μία χαρακτηριστική περίπτωση, εξετάζοντας το χαρακτηριστικό αυτό για δύο άντρες ομιλητές (με κωδικό 11 και 15) ηλικίας 26 και 25 ετών αντίστοιχα. Με τον τρόπο αυτό μειώνεται η εξάρτηση του F από το φύλο και την ηλικία. Αρχικά σχεδιάζουμε τα box plots των τιμών του F σε κάθε συναίσθημα για όλες τις προτάσεις, όπως φαίνεται στο σχήμα 5.11. Στο 4ο κανάλι δηλαδή στα 3000-5500Hz παρατηρούμε ότι ο ομιλητής 11 εκφράζει την πλήξη με υψηλότερες συχνότητες από τον

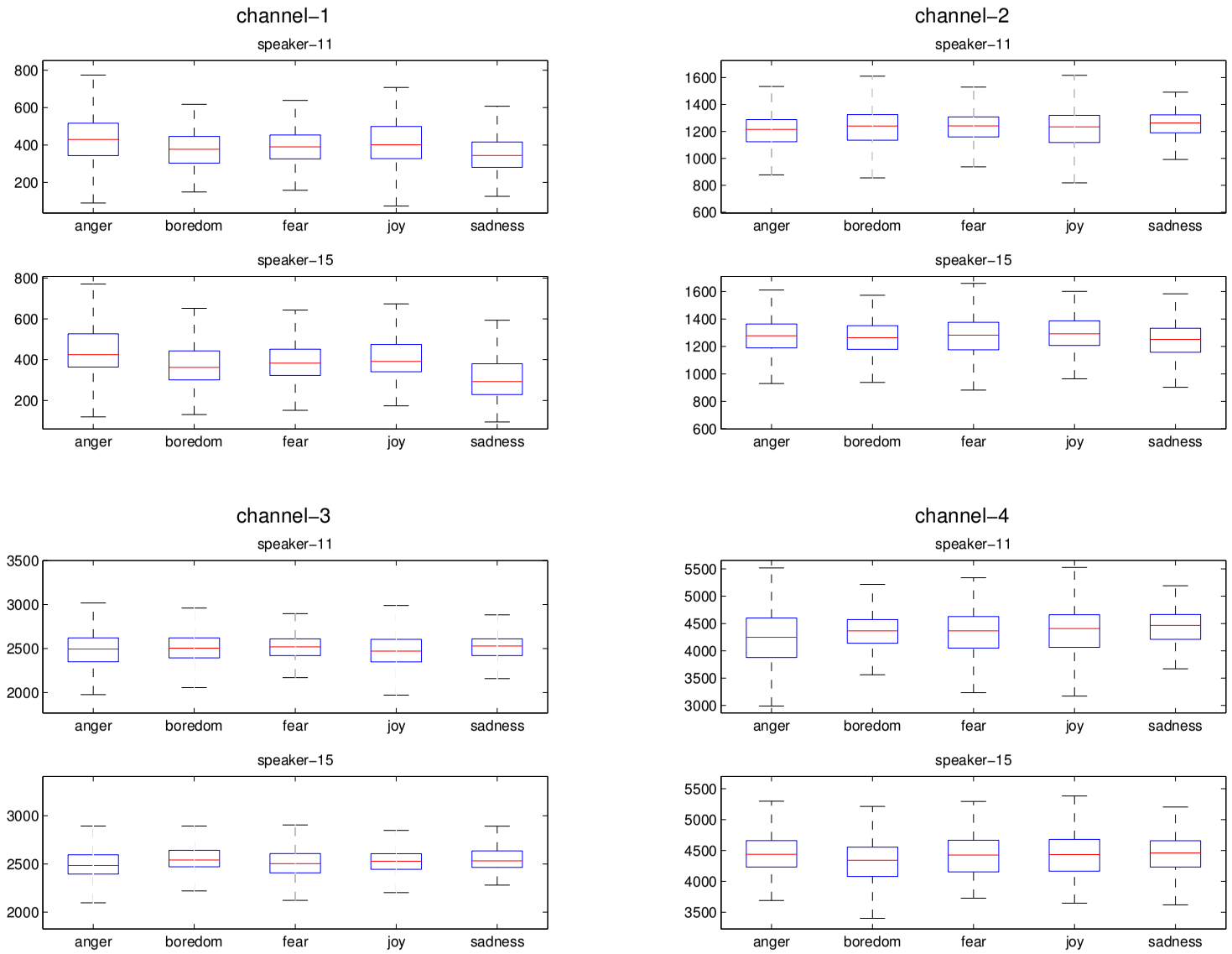


Σχήμα 5.9: Box plots των τιμών του σταθμισμένου μέσου όρου στιγμιαίας συχνότητας (F).

ομιλητή 15 και αντίστροφα ισχύει για το θυμό. Στο 1ο και 2ο κανάλι επίσης, το συναίσθημα της λύπης εκφράζεται με πιο υψηλές συχνότητες σε μικρότερο εύρος στον ομιλητή 11 και χαμηλότερες συχνότητες σε μεγαλύτερο εύρος στον ομιλητή 15.



Σχήμα 5.10: Γραφική παράσταση της πρότασης "Heute abend könnte ich es ihm sagen." ("Tonight I could tell him.") για το σταθμισμένο μέσο όρο στιγμιαίας συχνότητας (F) α) σε 4 προκαθορισμένα κανάλια και β) γύρω από τις συχνότητες των 4 πρώτων formants. Η οριζόντια γραμμή δείχνει το μέσο όρο των τιμών σε όλη τη χρονική διάρκεια του σήματος. Το F έχει υποστεί ομαλοποίηση, έτσι ώστε να μη λαμβάνονται υπόψη οι γρήγορες μεταβολές που οφείλονται στην παράγωγο του στιγμιαίου πλάτους και δεν περιέχουν χρήσιμη πληροφορία. Τα σήματα έχουν διαφορετική χρονική διάρκεια.

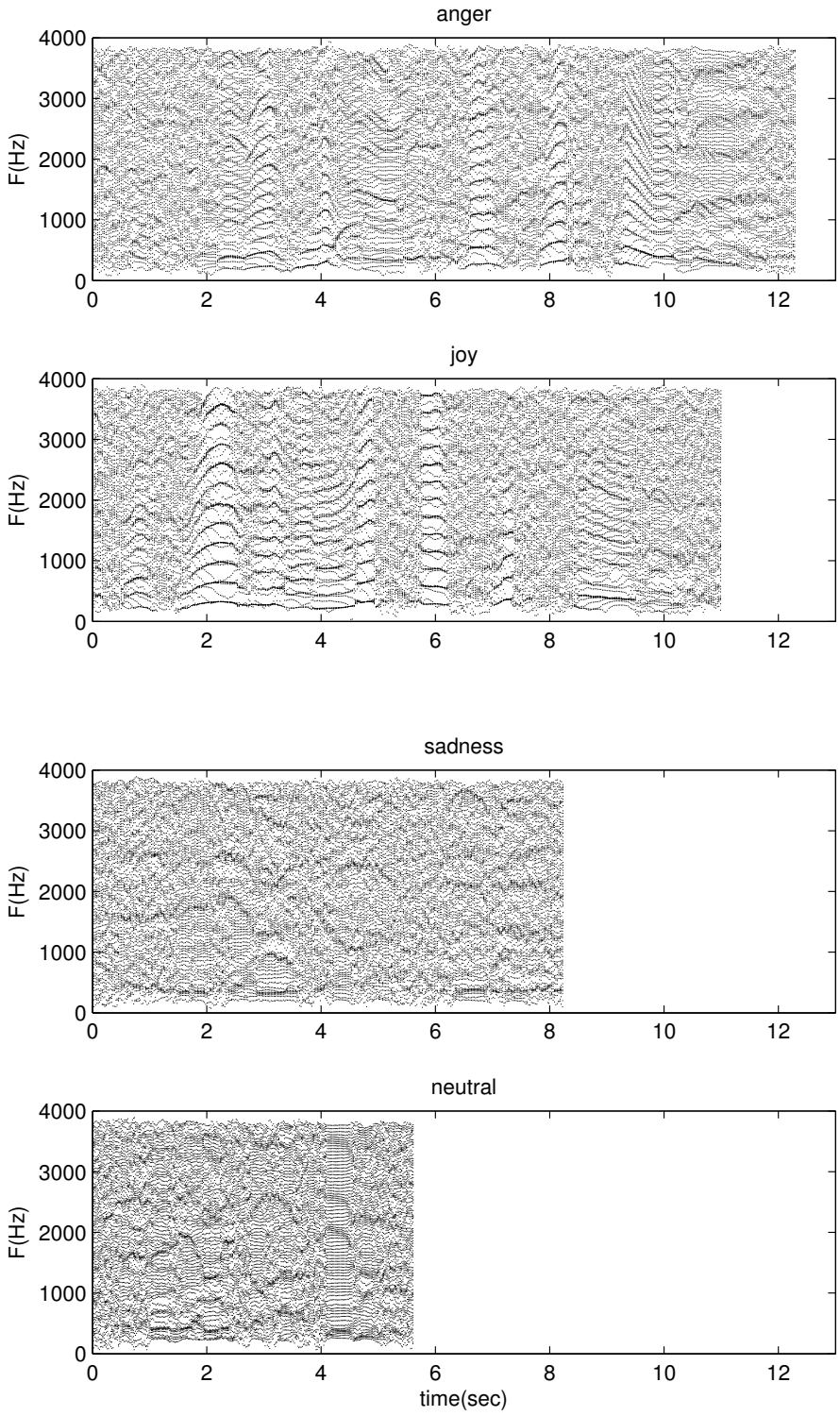


Σχήμα 5.11: Box plots των τιμών του σταθμισμένου μέσου όρου στιγμιαίας συχνότητας (F) για τους ομιλητές 11 και 15.

Ένα διάγραμμα που περιέχει μεγάλο όγκο πληροφορίας για το χαρακτηριστικό F είναι το πυκνόγραμμα. Το πυκνόγραμμα είναι μία αναπαράσταση του σήματος φωνής που έχει υποστεί ζωνοπερατό φιλτράρισμα και αποδιαμόρφωση. Η δομή αυτή χρησιμεύει στον εντοπισμό των διαμορφωτριών συχνοτήτων (formants), οι οποίες βρίσκονται στις πυκνές περιοχές του πυκνογράμματος. Το σήμα φωνής φιλτράρεται από 73 ζωνοπερατά Gabor φίλτρα, εύρους 400Hz το καθένα και ομοιόμορφα κατανεμημένα σε κέντρα συχνοτήτων από 0 έως 3800Hz με 50% επικάλυψη. Επίσης παραθυροποιείται με παράθυρα μήκους 10ms με 5ms επικάλυψη ανάμεσά τους.

Στο σχήμα 5.12 απεικονίζουμε τα πυκνογράμματα για έναν ομιλητή μίας πρότασης που έχει ειπωθεί με θυμό, χαρά, φόβο και το ουδέτερο συναίσθημα. Μεταξύ των συναισθημάτων παρατηρούνται αρκετές διαφορές. Στη χαρά είναι εμφανή τα formants των φωνηέντων, ιδιαίτερα του 'i' και του 'o'. Παρόμοια παρατήρηση, αλλά σε λίγο μικρότερη έκταση ισχύει και για το θυμό, όπου διακρίνονται επίσης οι θέσεις των formants. Στο συναίσθημα του θυμού εντοπίζονται καλύτερα απ΄ ότι στη χαρά τα formants του φωνήεντος 'e' στη λέξη abgeben. Στη λύπη και λιγότερο στο ουδέτερο δεν είναι τόσο εμφανείς οι θέσεις των formants. Αυτή η αντιθεση μεταξύ θυμού, χαράς και λύπης, ουδέτερου μπορεί να εξηγηθεί επειδή στο πρώτο ζευγάρι συναισθημάτων δίνεται έμφαση στα φωνήεντα και στις λέξεις κατά τη διάρκεια της εκφοράς. Αντίθετα, στο δεύτερο ζευγάρι ο λόγος είναι πιο ήπιος και υποτονικός.

Γενικά, ο σταθμισμένος μέσος όρος της στιγμιαίας συχνότητας ανακλά την τονικότητα του λόγου, αλλά ταυτόχρονα επηρεάζεται και από τον ομιλητή, για αυτό και όπως θα δούμε δίνει μέτρια αποτελέσματα αναγνώρισης.

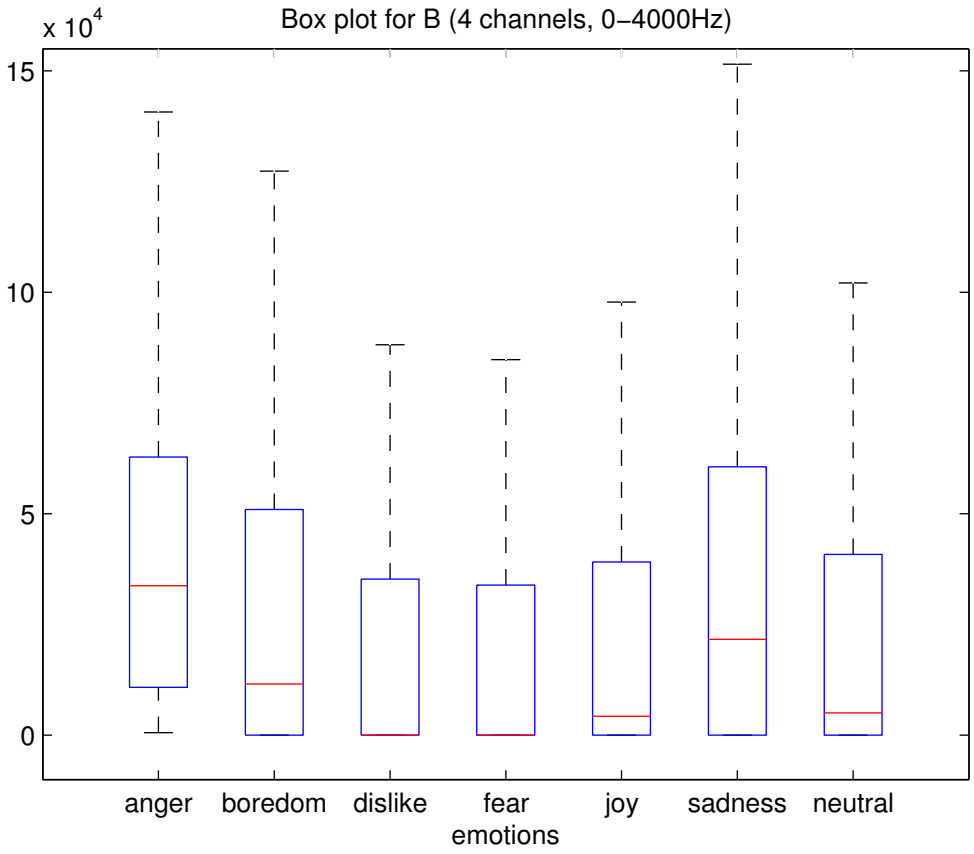


Σχήμα 5.12: Πυκνόγραμμα των τιμών της σταθμισμένης μέσης συχνότητας F υπολογισμένης σε 73 κανάλια πλάτους $400Hz$ ομοιόμορφα κατανεμημένων στις συχνότητες $0 - 4000Hz$. Μεταξύ των καναλιών υπάρχει 50% επικάλυψη. Τα σχήματα αφορούν την πρόταση a02: "Das will sie am Mittwoch abgeben." ("She will hand it in on Wednesday.") που έχει εκφωνηθεί από τον ομιλητή 03.

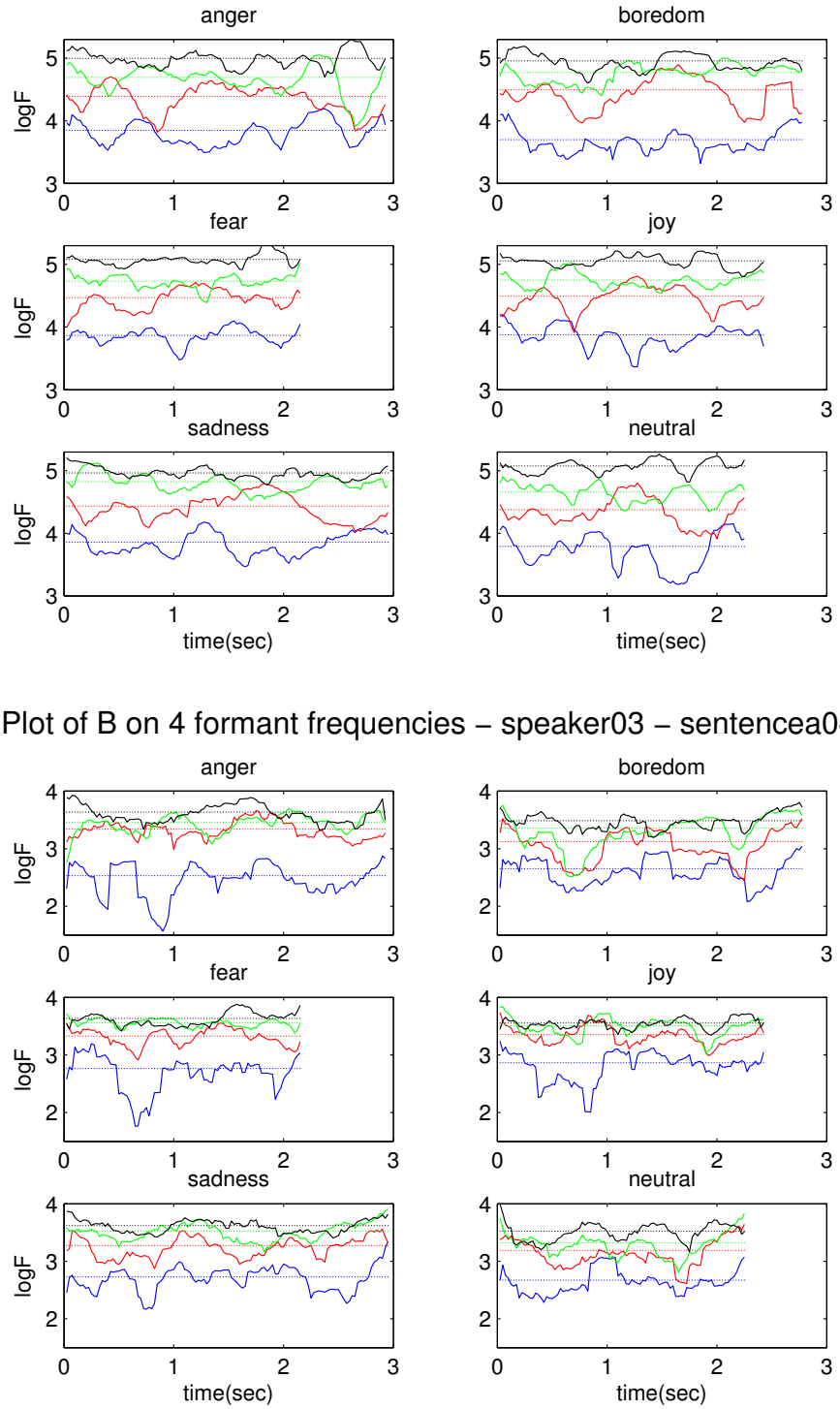
5.3.2 Σταθμισμένη Απόκλιση Στιγμιαίας Συχνότητας (B)

Εκτός από το μέσο όρο της στιγμιαίας συχνότητας, μπορούμε να υπολογίσουμε και τη σταθμισμένη απόκλιση στιγμιαίας συχνότητας, που αποτελεί μία εκτίμηση του εύρους των διαμορφωτριών συχνοτήτων (formants). Στο σχήμα 5.13 βλέπουμε τα box plots των τιμών του B για όλα τα συναισθήματα. Παρατηρούμε ότι ο θυμός έχει τις μεγαλύτερες τιμές του B και το μεγαλύτερο εύρος, πράγμα που υποδηλώνει ότι το συναίσθημα αυτό παρουσιάζει τη μεγαλύτερη μεταβλητότητα στο χώρο των συχνοτήτων. Αντίθετα, η απαρέσκεια, ο φόβος και το ουδέτερο εμφανίζουν μικρότερο εύρος τιμών με περιορισμένη μεταβλητότητα, οπότε ο λόγος στα συναισθήματα αυτά είναι πιο μονότονος.

Στο σχήμα 5.14 παρατηρούμε τη χρονική εξέλιξη του χαρακτηριστικού αυτού για μία τυχαία πρόταση και τυχαίο ομιλητή. Στις υψηλές συχνότητες επιβεβαιώνεται ότι ο φόβος έχει το μεγαλύτερο μέσο όρο μεταβλητότητας συχνοτήτων, ενώ απότομες μεταβολές στο μέγεθος B παρουσιάζουν ο θυμός και η πλήξη. Η λύπη σε αυτό το διάστημα συχνοτήτων έχει μικρή τιμή μέσου όρου του B και μικρή μεταβλητότητα. Παρόμοια πράγματα ισχύουν στις χαμηλές συχνότητες για το θυμό και το φόβο, ενώ η λύπη και το ουδέτερο έχουν μικρή απόκλιση από το μέσο όρο του B.



Σχήμα 5.13: Box plot των τιμών της σταθμισμένης απόκλισης στιγμιαίας συχνότητας (B).



Σχήμα 5.14: Γραφική παράσταση της πρότασης "Heute abend konnte ich es ihm sagen." ("Tonight I could tell him.") για τη σταθμισμένη απόκλιση στιγμιαίας συχνότητας (B) α) σε 4 προκαθορισμένα κανάλια και β) γύρω από τις συχνότητες των 4 πρώτων formants. Η οριζόντια γραμμή δείχνει το μέσο όρο των τιμών σε όλη τη χρονική διάρκεια του σήματος. Το B έχει υποστεί ομαλοποίηση, έτσι ώστε να μη λαμβάνονται υπόψη οι γρήγορες μεταβολές που οφείλονται στην παράγωγο του στιγμιαίου πλάτους και δεν περιέχουν χρήσιμη πληροφορία. Τα σήματα έχουν διαφορετική χρονική διάρκεια.

5.4 TEO-Auto-Env

Τα AM-FM χαρακτηριστικά διαμόρφωσης είναι αρκετά θορυβώδη. Για το λόγο αυτό και για να διατηρήσουμε την πληροφορία που παίρνουμε από το στιγμιαίο πλάτος, ομαλοποιούμε το μέγεθος αυτό και στη συνέχεια υπολογίζουμε το εμβαδόν κάτω από την ομαλοποιημένη καμπύλη. Έπειτα, για να εξαλειφθούν οι απότομες μεταβολές, μπορεί να υπολογίζουμε την αυτοσυσχέτιση του σήματος αυτού.

Το χαρακτηριστικό που προκύπτει ονομάζεται TEO-Auto-Env και έχει χρησιμοποιηθεί για την αναγνώριση του άγχους στη φωνή. Το πλεονέκτημά του είναι ότι μειώνονται οι γρήγορες διακυμάνσεις του σήματος φωνής, αλλά παρόλα αυτά διατηρούνται οι αλλαγές στη φωνή που οφείλονται στο άγχος. Με τον τρόπο αυτό απομακρύνεται η επιροή του λεκτικού περιεχομένου στη φωνή και δίνεται έμφαση στα μη γραμμικά χαρακτηριστικά διέγερσης (excitation) [99]. Όπως θα δούμε και στη συνέχεια με βάση το χαρακτηριστικό αυτό μπορούμε να ξεχωρίσουμε τα συναισθήματα σε αρκετά μεγάλο βαθμό.

Η αυτοσυσχέτιση ενός σήματος πεπερασμένου μήκους $x(n)$ ορίζεται ως

$$R_n(k) = \sum_{m=-\infty}^{\infty} x(m)x(m+k)$$

Λόγω του πεπερασμένου μήκους του σήματος, η αυτοσυσχέτιση φθίνει με την πάροδο του χρόνου, γιατί όσο αυξάνεται ο χρονικός δείκτης k, εμπλέκονται όλο και λιγότερα δείγματα στον υπολογισμό της αυτοσυσχέτισης. Για να αποφευχθεί αυτή η μείωση [50] (κεφάλαιο 4), στο τρέχον frame που ορίζεται από το παράθυρο w_1 μπορούν να προστεθούν δείγματα και από το επόμενο frame που ορίζεται από το w_2. Δηλαδή ορίζουμε τα w_1 και w_2 ως εξής:

$$w_1(m) = \left\{ \begin{array}{c} 1, \;\; 0 \leq m \leq N-1 \\ 0, \;\; \text{αλλιώς} \end{array} \right\}$$

$$w_2(m) = \left\{ \begin{array}{c} 1, \;\; 0 \leq m \leq N-1+K \\ 0, \;\; \text{αλλιώς} \end{array} \right\}$$

όπου το K είναι η μέγιστη καθυστέρηση φάσης. Με τον τρόπο αυτό η νέα τροποποιημένη συνάρτηση αυτοσυσχέτισης γίνεται:

$$\hat{R}_n(k) = \sum_{m=0}^{N-1} x(n+m)x(n+m+k), \quad 0 \leq k \leq K$$

Η τροποποιημένη συνάρτηση αυτοσυσχέτισης $\hat{R}(k)$ έχει κορυφές στα πολλαπλάσια της περιόδου ενός περιοδικού σήματος, όπως και η αρχική συνάρτηση $R(k)$, αλλά δεν παρουσιάζει φθίνουσα πορεία στο τέλος κάθε frame.

Στα σχήματα 5.15 και 5.16 απεικονίζουμε γραφικά το αρχικό σήμα του φωνήεντος "a" για 6 συναισθήματα, το ζωνοπερατό σήμα, το στιγμιαίο πλάτος αυτού και την αυτοσυσχέτιση του πλάτους. Μετά τον υπολογισμό του στιγμιαίου πλάτους, έχουμε χωρίσει το σήμα σε frames διάρκειας 10msec και με επικάλυψη 5msec. Παρατηρούμε ότι η περιοδικότητα του στιγμιαίου πλάτους διαφέρει ανάμεσα στα συναισθήματα. Ο θυμός, η πλήξη και η λύπη έχουν μικρότερη περίοδο από το φόβο, τη χαρά και το ουδέτερο. Επιπλέον, το στιγμαίο πλάτος εμφανίζει μικρές τιμές στο θυμό και στη χαρά, ενώ μεγαλύτερες τιμές στα υπόλοιπα συναισθήματα, πράγμα που αντικατοπτρίζεται και στην αυτοσυσχέτιση.

Εκτός από τα κλασσικά μεγέθη του μέσου όρου και της τυπική απόκλισης που υπολογίζονται συνήθως, μπορεί να βρεθεί το εμβαδόν κάτω από την καλμπύλη του στιγμαίου πλάτους και της αυτοσυσχέτισής του. Στο σχήμα 5.17 φαίνονται τα box plots του εμβαδού των δύο μεγεθών. Παρατηρείται μεγάλη διακριτότητα ανάμεσα στα συναισθήματα με βάση τα χαρακτηριστικά αυτά. Θα μπορούσαμε ίσως να υποθέσουμε ότι το εμβαδόν του στιγμιαίου πλάτους θα είναι λίγο πιο αποτελεσματικό από το εμβαδόν της αυτοσυσχέτισης, γιατί υπάρχει λιγότερη επικάλυψη των τιμών του ανάμεσα στα συναισθήματα.

Στο σχήμα 5.18 απεικονίζονται τα box plots του TEO-Auto-Env σε 4 κανάλια. Παρατηρούμε ότι όσο αυξάνεται η συχνότητα, οι τιμές του TEO-Auto-Env φθίνουν. Στο 1ο κανάλι των χαμηλών συχνοτήτων, ο θυμός έχει τις μικρότερες τιμές και το μικρότερο εύρος τιμών. Στη συνέχεια ακολουθεί η χαρά και μετά η λύπη και η πλήξη. Η αποστροφή, ο φόβος και το ουδέτερο παρουσιάζουν μεγαλύτερες τιμές του Auto-Env στο 1ο κανάλι, καθώς και μεγαλύτερο έυρος τιμών. Χρήσιμες πληροφορίες παίρνουμε και από το 2ο κανάλι, όπου το πιο διακριτό συναίσθημα είναι η χαρά, που έχει υψηλές τιμές του χαρακτηριστικού αυτού συγκεντρωμένες σε στενό εύρος. Τέλος, η λύπη είναι το συναίσθημα που διαφοροποιείται σε σχέση με τα υπόλοιπα στις υψηλές συχνότητες, αφού εκεί παρουσιάζει μεγαλύτερες τιμές από τα άλλα συναισθήματα.

Απεικονιζουμε το χαρακτηριστικό TEO-Auto-Env με τη βοήθεια φασματογραμμάτων (spectrograms) όπου ο οριζόντιος άξονας κάθε διαγράμματος αντιστοιχεί στη χρονική εξέλιξη του σήματος και ο κατακόρυφος άξονας στα κανάλια των συχνοτήτων. Επειδή στα box plots παρατηρήθηκε ότι οι τιμές του TEO-Auto-Env στα τελευταία κανάλια είναι πολύ μικρές, για τα φασματογραφήματα υπολογίσαμε το χαρακτηριστικό αυτό σε συχνότητες 0-4000 Hz με 13 κανάλια. Κάθε φασματογράφημα κατασκευάζεται για μία πρόταση, οπότε στο ίδιο σχήμα τοποθετούμε την ίδια πρόταση και τον ίδιο ομιλητή για όλα τα διαφορετικά συναισθήματα με τα οποία ο ομιλητής έχει εκφράσει αυτήν την πρόταση. Με τον τρόπο αυτό προέκυψαν 100 διαφορετικά σχήματα. Στο σχήμα 5.19 βλέπουμε δύο αντιπροσωπευτικά φασματογραφήματα και αναφέρουμε ορισμένες ομοιότητες που παρατηρήθηκαν στην πλειονότητά όλων των σχημάτων.

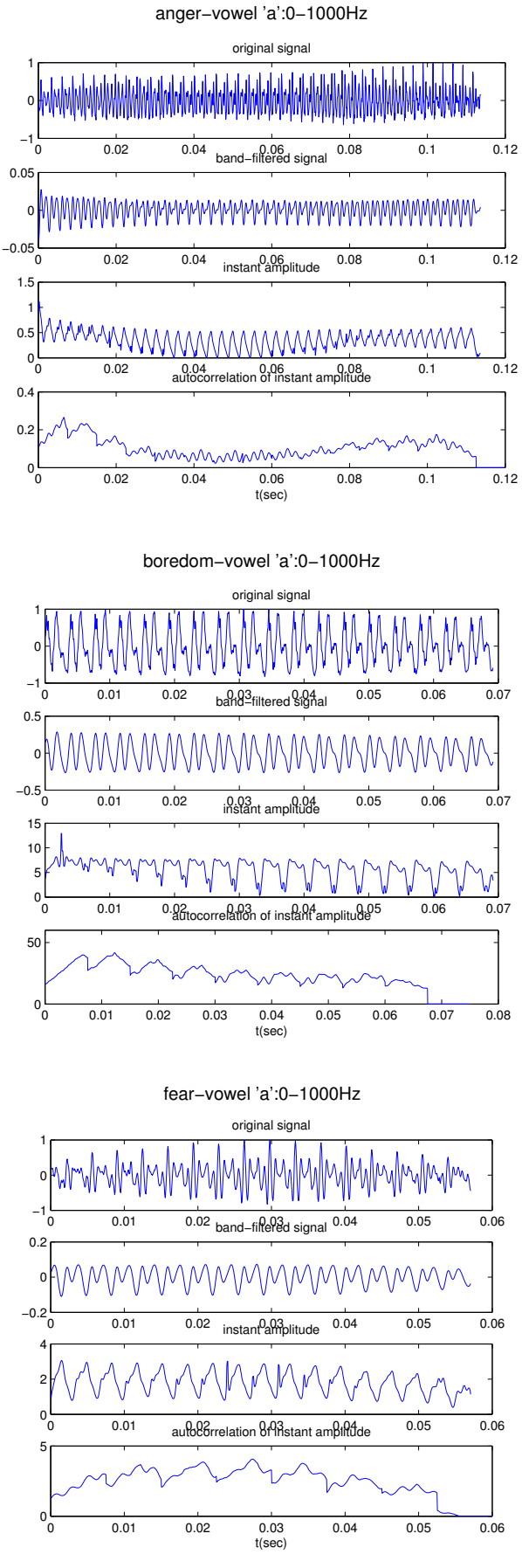
θυμός: Το TEO-Auto-Env για το συναίσθημα αυτό είναι περισσότερο εκτεταμένο στον άξονα του χρόνου και των συχνοτήτων σε σχέση με κάθε άλλο συναίσθημα. Παρόλα αυτά οι τιμές του χαρακτηριστικού αυτού δεν είναι τόσο μεγάλες όσο σε άλλα συναισθήματα. Η παρατήρηση αυτή συμφωνεί με την παρατήρηση από τα box plots, όπου το συναίσθημα του θυμού είχε μικρότερη διάμεσο και 25%, 75% τιμή εκατοστημορίου από τα υπόλοιπα συναισθήματα.

πλήξη: Η πλήξη εκτείνεται συνήθως στις συχνότητες 0-2000 Hz και εμφανίζει τις περισσότερες φορές μεγάλες τιμές στις μικρές συχνότητες, πράγμα που σημαίνει ότι το συναίσθημα αυτό εκφράζεται με σχετικά βαθιά φωνή.

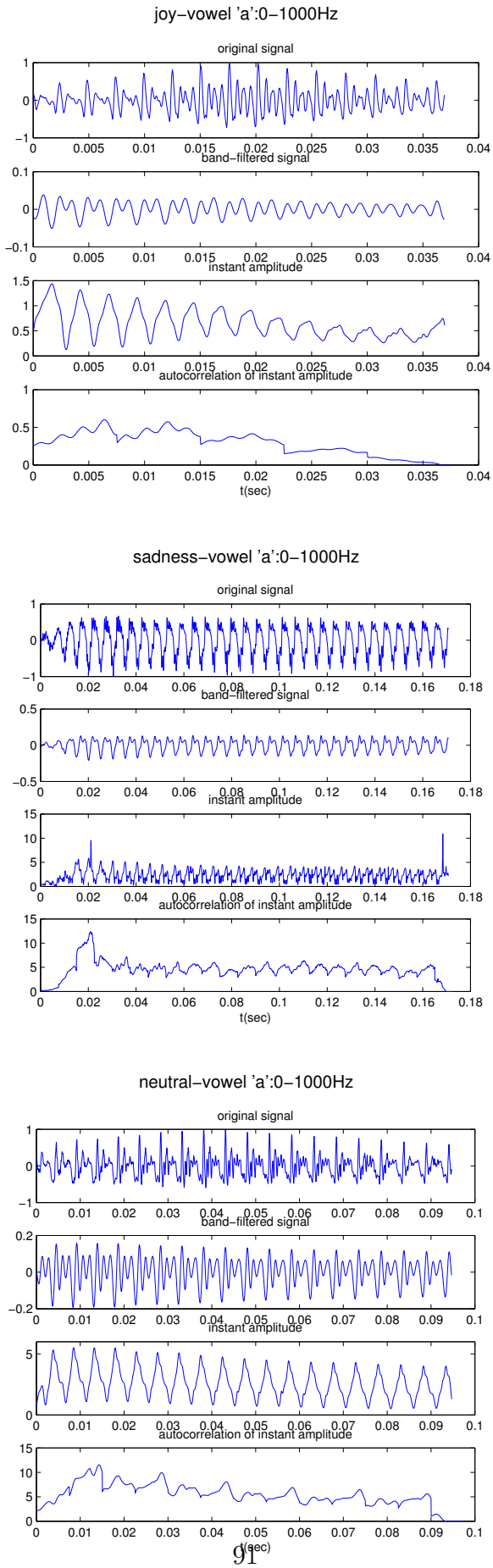
αποστροφή: Το συναίσθημα της αποστροφής δεν έχει κάποιο πολύ φανερό κοινό χαρακτηριστικό σε όλες τις προτάσεις. Παρόλα αυτά συνήθως κυμαίνεται από 0-1500 Hz και παρατηρείται μετακίνηση μεγίστων των τιμών του Auto-Env προς τα δεξιά. Αυτό φανερώνει ότι η αποστροφή μπορεί να εκφραστεί με μεγαλύτερη ένταση στο τέλος της πρότασης.

φόβος: Ο φόβος δεν έχει επίσης κάποια συγκεκριμένη μορφή στα φασματογραφήματα. Τις περισσότερες φορές εκτείνεται στις χαμηλές συχότητες 0-2000 Hz και μπορεί να εμφανίζει μεγαλύτερες τιμές στη μέση ή στο τέλος της πρότασης.

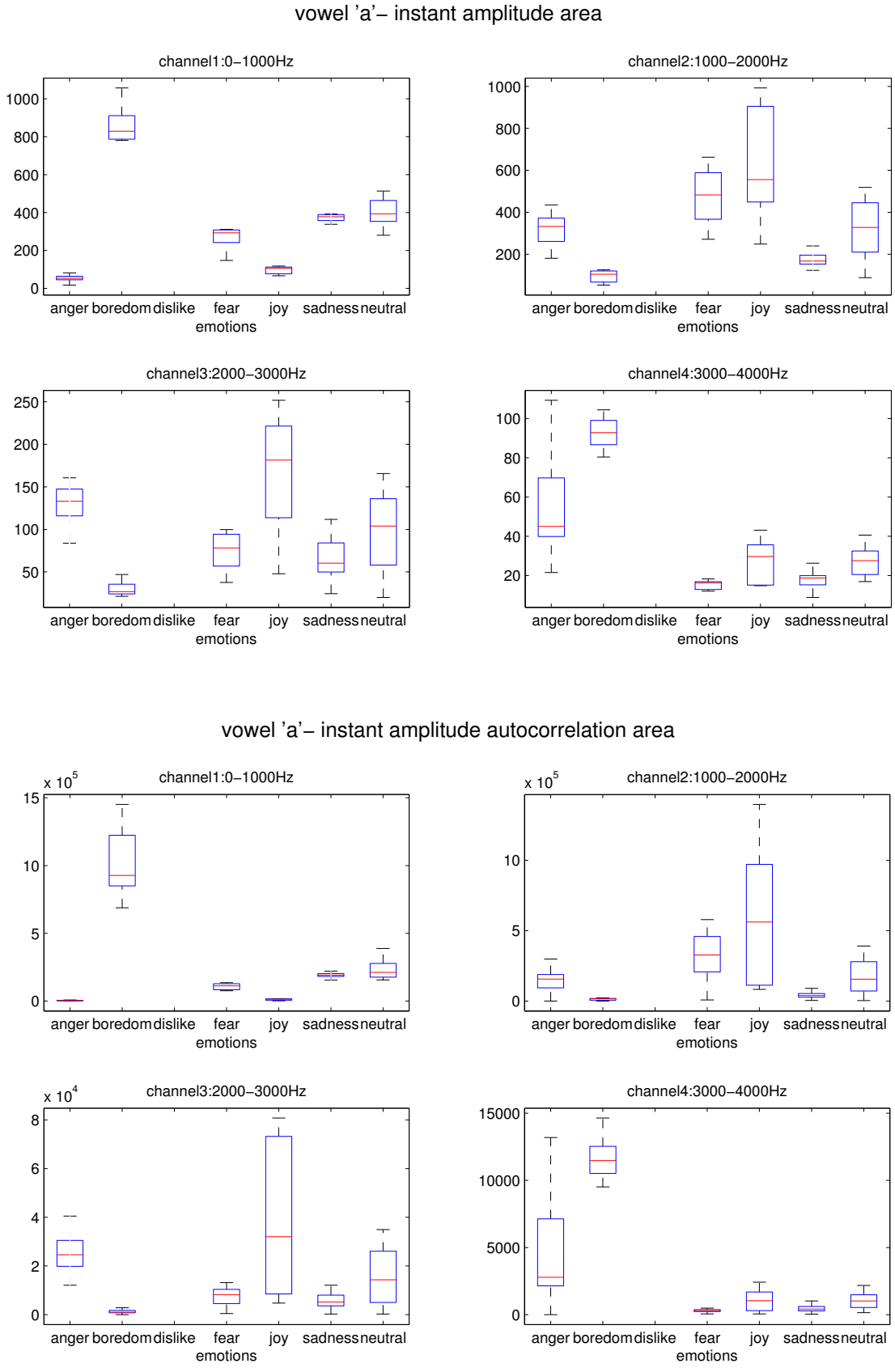
χαρά: Η χαρά είναι συναίσθημα αρκετά εκτεταμένο στο χώρο των συχνοτήτων (συνήθως για μία δεδομένη πρόταση, πιο πολύ από το φόβο και λιγότερο από το θυμό). Επίσης,



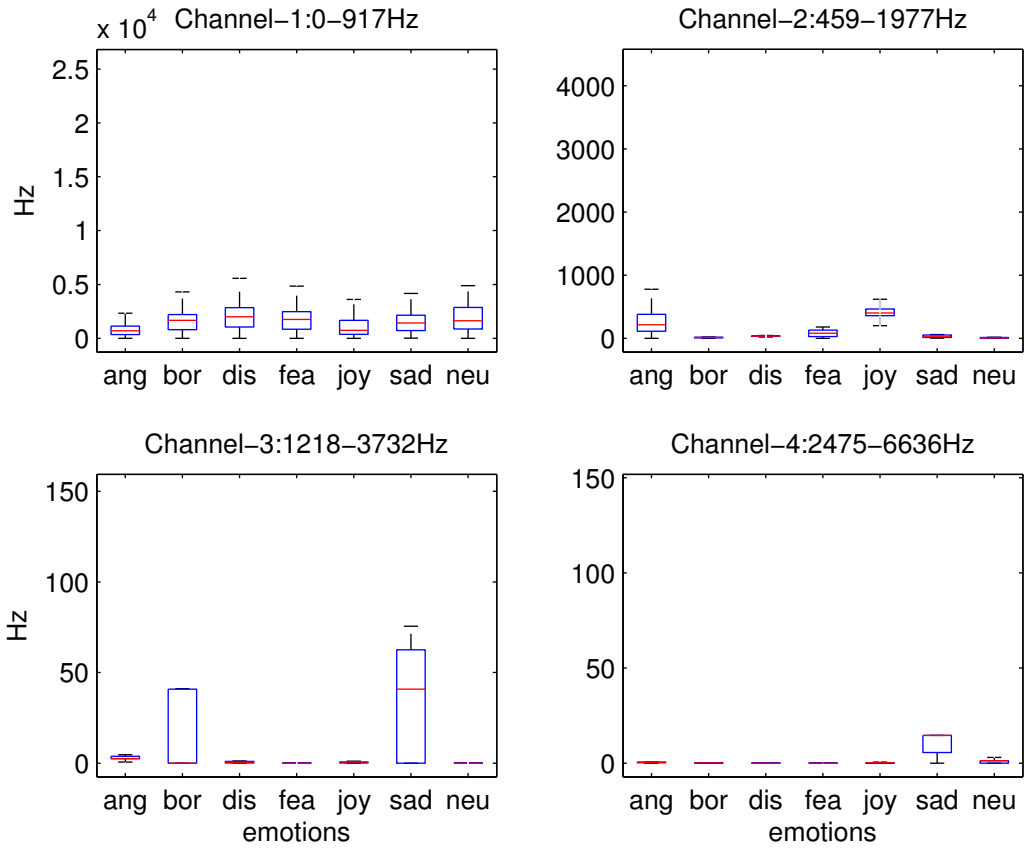
Σχήμα 5.15: Γραφική απεικόνιση του φωνήεντος 'a', του ζωνοπερατού σήματος, του στιγμιαίου πλάτους και της αυτοσυσχέτισης του πλάτους στο θυμό, την πλήξη και το φόβο.



Σχήμα 5.16: Γραφική απεικόνιση του φωνήεντος 'a', του ζωνοπερατού σήματος, του στιγμιαίου πλάτους και της αυτοσυσχέτισης του πλάτους στη χαρά, τη λύπη και το ουδέτερο.



Σχήμα 5.17: Box plots του εμβαδού του στιγμαίου πλάτους και της αυτοσυσχέτισης του στιγμιαίου πλάτους του φωνήεντος 'a' για 69 συναισθήματα σε 4 κανάλια.



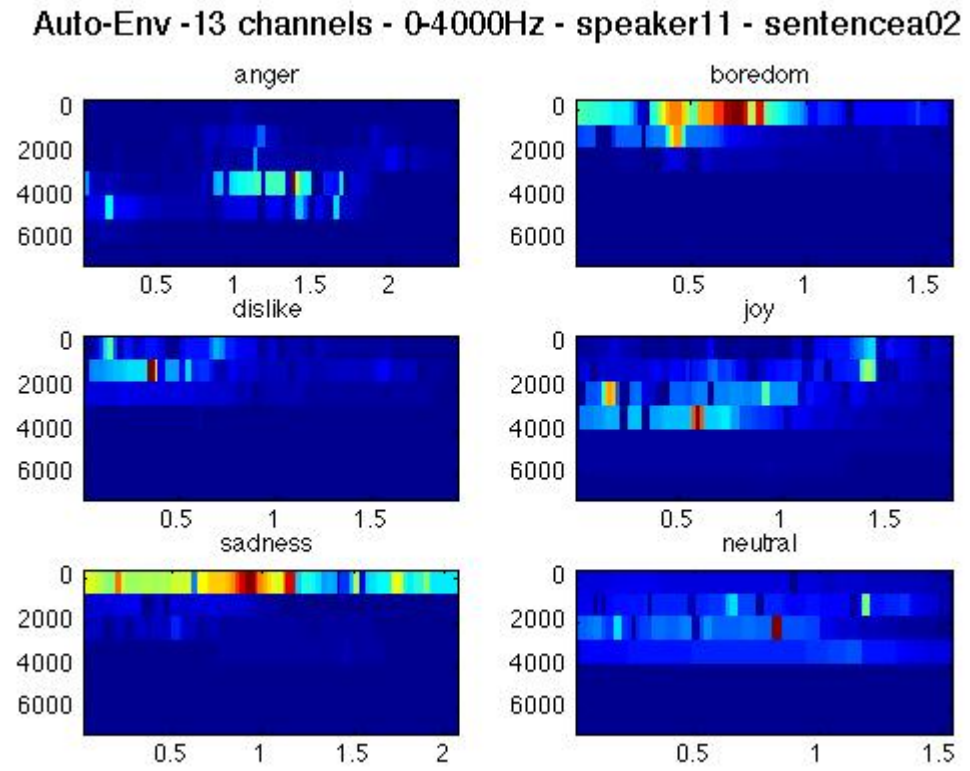
Σχήμα 5.18: Box plots των τιμών του TEO-Auto-Env σε 4 κανάλια για όλες τις προτάσεις της βάσης δεδομένων Berlin Database of Emotional Speech.

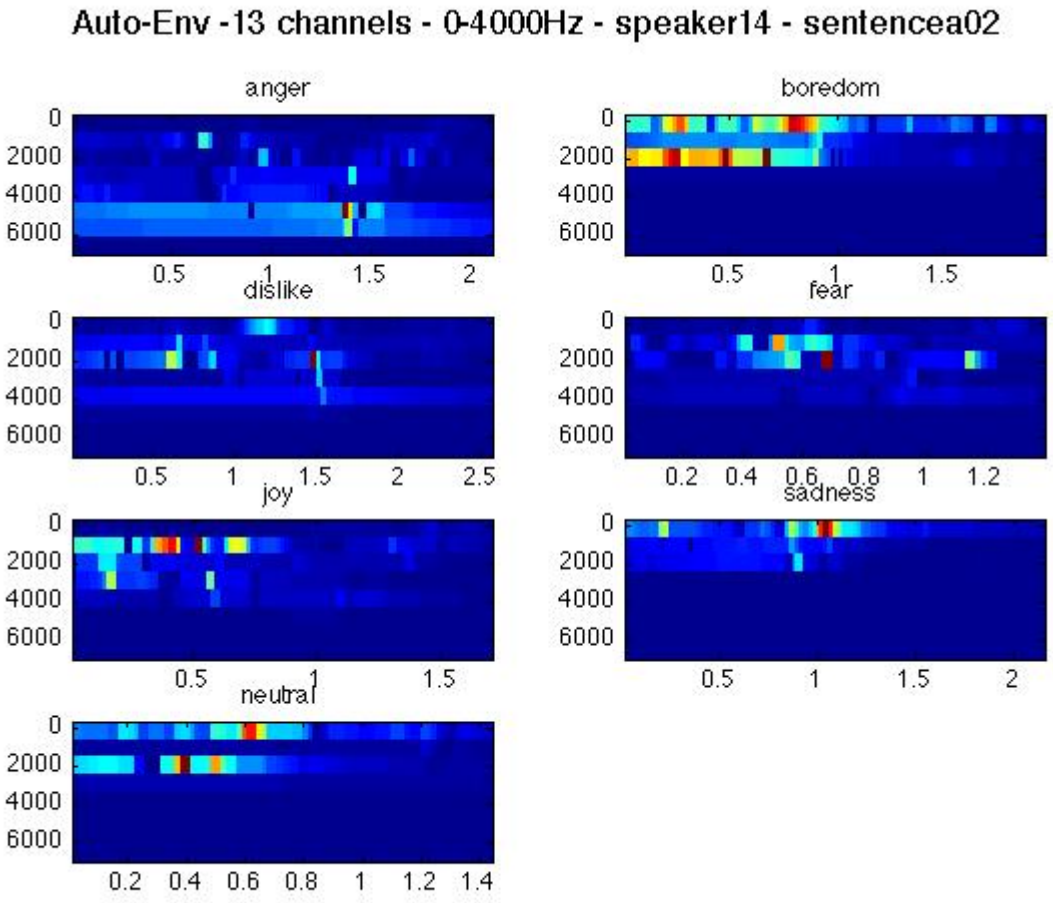
παρουσιάζει μεγάλες τιμές ακανόνιστα στο χρόνο, δηλαδή στην αρχή, στη μέση είτε και στο τέλος μίας πρότασης. Αυτό σήμαινει ότι για το συναίσθημα της χαράς υπάρχουν πολλοί τρόποι έκφρασης, με ψιλή φωνή συνήθως σε διάσπαρτα μέρη της πρότασης.

λύπη: Το συναίσθημα της λύπης έχει σχεδόν σε όλες τις προτάσεις συγκεκριμένη μορφή: παρουσιάζει μεγάλες τιμές στις χαμηλές συχνότητες και σε στενό εύρος αυτών, συνήθως στα 0-500 Hz. Όσον αφορά τη χρονική εξέλιξη, έμφαση δίνεται συνήθως στην αρχή της πρότασης, αλλά μπορεί να δοθεί και σε άλλα σημεια.

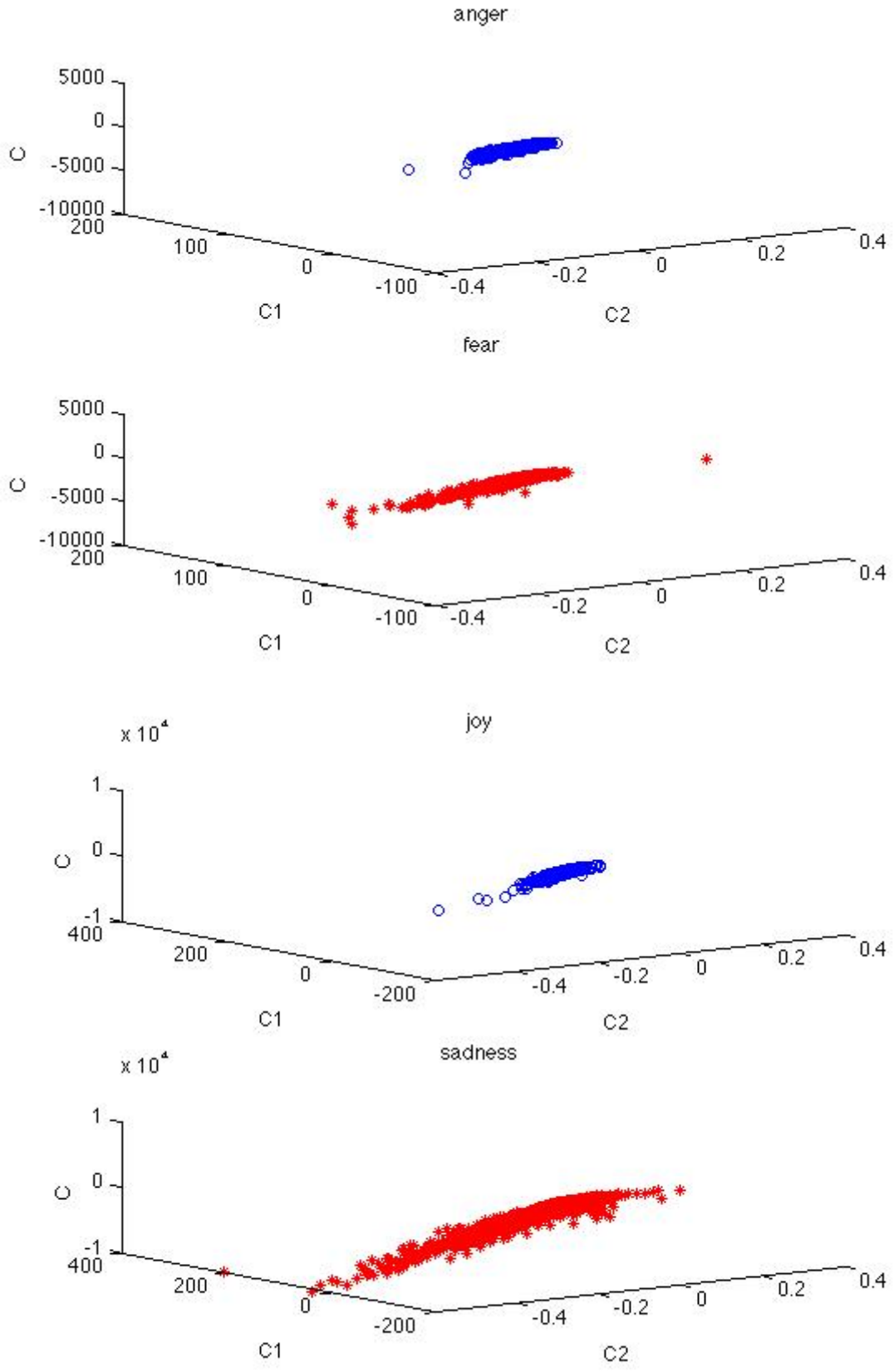
ουδέτερο: Το ουδέτερο συγκεντρώνει το χαρακτηριστικό TEO-Auto-Env στα 0-2000 Hz. Εμφανίζει επίσης φθίνουσα πορεία στο χρόνο, πράγμα που έρχεται σε συμφωνία με τη φθίνουσα πορεία του pitch, που διαπιστώθηκε σε προηγούμενη έρευνα [59] για το ουδέτερο.

Τέλος, στο σχήμα 5.20 απεικονίζονται οι συντελεστές C_0, C_1 και C_2 του πολυωνύμου $p(X)$ σε 3 διαστάσεις για τα αντίθετα ζευγάρια συναισθημάτων χαρά - λύπη και θυμός - φόβος. Παρατηρούμε ότι υπάρχει επικάλυψη μεταξύ των τρισδιάστατων σημείων για τα συναισθήματα. Παρόλα αυτά η λύπη διαφοροποιείται από τη χαρά, επειδή παρουσιάζει μεγαλύτερη διασπορά και παρόμοια ισχύει για το φόβο σε σχέση με το θυμό. Τα συναισθήματα όπως η χαρά και ο θυμός, στα οποία ο ομιλιτής εξωτερικεύει την ψυχολογική του κατάσταση εμφανίζουν μικρότερη διασπορά στις τιμές των C_0, C_1 και C_2 σε σχέση με τη λύπη και το φόβο, τα οποία ο ομιλητής βιώνει εσωτερικά.





Σχήμα 5.19: Φασματογράφημα των τιμών του TEO-Auto-Env υπολογισμένο σε 13 κανάλια για την πρόταση "Das will sie am Mittwoch abgeben." ("She will hand it in on Wednesday.") εκφρασμένη από τους ομιλητές 11 και 14.

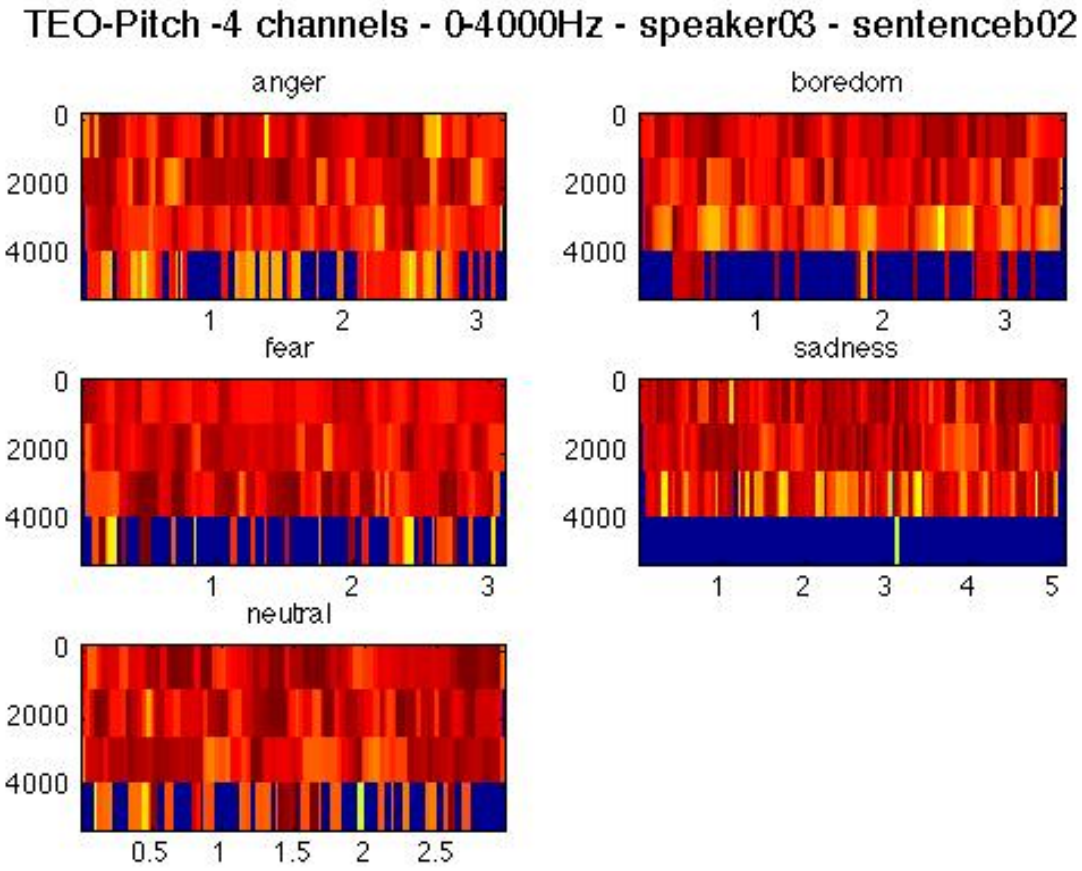


Σχήμα 5.20: Τρισδιάστατη αναπαράσταση των συντελεστών C_0, C_1 και C_2 του πολυωνύμου $p(X)$ για τα ζευγάρια συναισθημάτων θυμός - φόβος και χαρά - λύπη

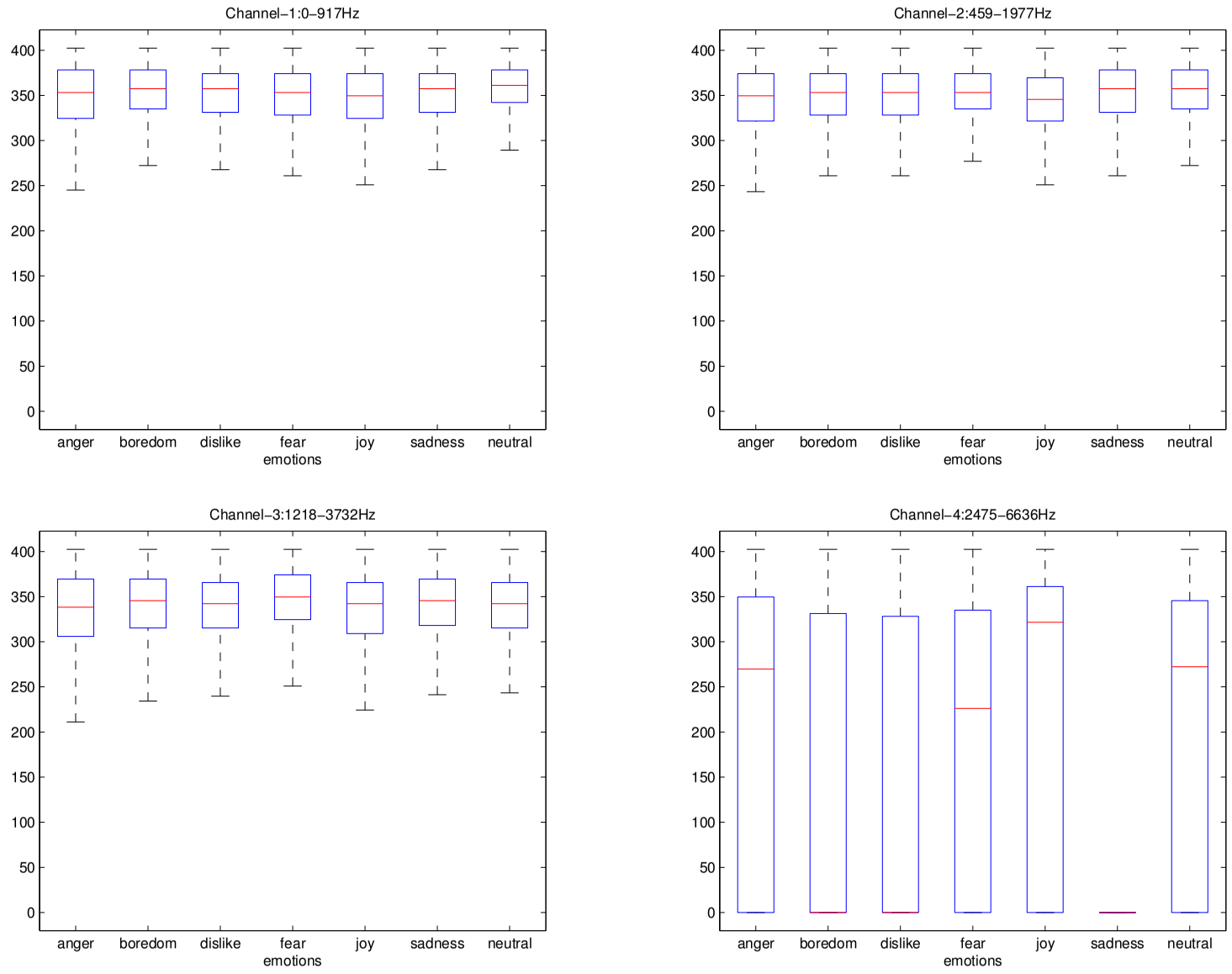
5.5 TEO-Pitch

Το TEO-Pitch έχει χρησιμοποιηθεί στην αναγνώριση των ειδών του άγχους [98]. Το χαρακτηριστικό αυτό αποτελεί μία εκτίμηση της περιοδικότητας των ESA μεγεθών.

Στο σχήμα 5.22 βλέπουμε το box plot του TEO-Pitch που έχει υπολογιστεί σε 4 κανάλια. Παρατηρούμε ότι οι μεγαλύτερες διαφορές μεταξύ των συναισθημάτων φαίνονται στο 4ο κανάλι για το χαρακτηριστικό αυτό.Στο κανάλι αυτό το TEO-Pitch για τη λύπη έχει σχεδόν μηδενικές τιμές, ενώ η διάμεσος της πλήξης και της αποστροφής στο ίδιο κανάλι είναι επίσης μηδενικές. Μεγαλύτερη τιμή διαμέσου έχει η χαρά, σχετικά μεγάλη τιμή έχουν ο θυμός και το ουδέτερο, ενώ μεσαία τιμή ο φόβος. Οι παρατηρήσεις αυτές φαίνονται και στα φασματογράμματα των προτάσεων. Ενδεικτικά παραθέτουμε το φασματόγραμμα της πρότασης b02 από τον ομιλητή 3 στο σχήμα 5.21. Η μεγάλη διαφορά του TEO-Pitch στις υψηλές συχνότητες μπορεί να σημαίνει ότι το συναίσθημα επηρεάζει την περιοδικότητα του σήματος στιγμιαίας συχνότητας σε υψηλές συχνότητες, και άρα ότι οι φωνητικές χορδές (που ρυθμίζουν κατά κύριο λόγο την περιοδικότητα της φωνής) ανοίγουν και κλείνουν με διαφορετικό ρυθμό ανάλογα με το συναίσθημα κατά την παραγωγή υψηλών συχνοτήτων.



Σχήμα 5.21: Φασματόγραμμα του χαρακτηριστικού TEO-Pitch για την πρόταση "Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter." ("They just carried it upstairs and now they are going down again.")



Σχήμα 5.22: Box plots των τιμών του TEO-Pitch για 4 κανάλια.

5.6 Περαιτέρω Επεξεργασία των Χαρακτηριστικών

Στην αναγνώριση συναισθήματος υπάρχει πολλή πλεονάζουσα πληροφορία στο σήμα της φωνής, που μεταφέρεται και στα χαρακτηριστικά που αναφέραμε. Για το λόγο αυτό και επειδή τα AM-FM χαρακτηριστικά διαμόρφωσης είναι θορυβώδη, μπορούμε να ομαλοποιούμε τις τιμές τους, ώστε να εξαλειφθούν οι απότομες μεταβολές, που δεν προσφέρουν κάποια χρήσιμη πληροφορία. Εφαρμόζουμε δύο διαδοχικές ομαλοποιήσεις:

- **Εξάλειψη NaNs και median φιλτράρισμα**
  Από τις διαιρέσεις με μηδέν, προκύπτουν τιμές που δεν ορίζονται (NaNs). Αυτές αντικαθίστανται με το μέσο όρο των w_1 γειτονικών τιμών που δεν ισούνται με (NaNs). Αν x_0 το σημείο του αρχικού σήματος για το οποίο $s(x_0) = NaN$, τότε από την παραπάνω αντικατάσταση το νέο σήμα που προκύπτει είναι:

$$y_1(x_0) = \sum_{i=x_0-w}^{x_0+w} s(i), \;\; s(i) \neq NaN, \;\; 2w+1 = w_1$$

  Στη συνέχεια εφαρμόζεται ένα median φιλτράρισμα, όπου η τιμή σε κάθε σημείο αντικαθίσταται με τη διάμεσο των τιμών του ίδιου σημείου και των δύο γειτονικών του, δηλαδή

$$y_2(i) = \left\{ \begin{array}{c} median(j, j = i-1, i, i+1), \;\; i = 2, \ldots, L-1 \\ \min(y_1(1), y_1(2)) \\ \min(y_1(L-1), y_1(L)) \end{array} \right\}$$

  όπου L το μήκος του αρχικού σήματος

- **Ομαλοποίηση μετακινούμενου μέσου όρου**
  Μετά από την εξάλειψη των NaN τιμών και το median φιλτράρισμα, εφαρμόζουμε ομαλοποίηση μέσου όρου, ώστε να πάρουμε τη γενική τάση μεταβολής του κάθε χαρακτηριστικού. Η τιμή κάθε σημείου γίνεται ίση με το μέσο όρο των γειτονικών του σημείων που ανήκουν σε ένα παράθυρο μήκους w_2. Στα άκρα του σήματος εφαρμόζεται ένας μέσος όρος από σημεία μεταβαλλόμενου μήκους. Το σήμα y_3 που προκύπτει από την ομαλοποίηση αυτή είναι:

$$y_3(i) = \left\{ \begin{array}{c} \frac{\sum_{j=1}^{2i-1} y_2(i)}{2i-1}, \;\; i = 1, \ldots, \lceil \frac{w_2}{2} \rceil - 1 \\ \frac{\sum_{j=i-\lfloor \frac{w_2}{2} \rfloor}^{j=i+\lfloor \frac{w_2}{2} \rfloor} y_2(i)}{w_2}, \;\; i = \lceil \frac{w_2}{2} \rceil, \ldots, L - \lceil \frac{w_2}{2} \rceil + 1 \\ \frac{\sum_{j=L-2i+2}^{L} y_2(i)}{2i-1}, \;\; i = L - \lceil \frac{w_2}{2} \rceil + 2, \ldots, L \end{array} \right\}, \quad w_2 \text{ περιττός}$$

$$y_3(i) = \left\{ \begin{array}{c} \frac{\sum_{j=1}^{2i-1} y_2(i)}{2i-1}, \;\; i = 1, \ldots, \frac{w_2}{2} - 1 \\ \frac{\sum_{j=i-\frac{w_2}{2}}^{j=i+\frac{w_2}{2}} y_2(i)}{w_2}, \;\; i = \frac{w_2}{2}, \ldots, L - \frac{w_2}{2} \\ \frac{\sum_{j=L-2i+2}^{L} y_2(i)}{2i-1}, \;\; i = L - \frac{w_2}{2} + 2, \ldots, L \end{array} \right\}, \quad w_2 \text{ άρτιος}$$

5.7 Συμπεράσματα

Στη συνέχεια, παραθέτουμε ορισμένα συμπεράσματα που προκύπτουν από τη μελέτη των χαρακτηριστικών και των γραφικών τους παραστάσεων, τα οποία θα είναι χρήσιμα για τη μετέπειτα ταξινόμηση των συναισθημάτων.

- Το χαρακτηριστικό της *θεμελιώδους συχνότητας* είναι το πιο διαδεδομένο και ταυτόχρονα αποτελεσματικό χαρακτηριστικό για την αναγνώριση συναισθήματος. Πιστεύουμε ότι σε συνεργασία με τα AM-FM χαρακτηριστικά διαμόρφωσης μπορεί να δώσει καλά αποτελέσματα ταξινόμησης.

- Ο *μέσος όρος της Teager ενέργειας* έχει διαφορετική κατανομή για τα συναισθήματα στο διάστημα συχνοτήτων 0-6000Hz. Κάποια συναισθήματα όπως ο θυμός έχουν μεγάλο φασματικό περιεχόμενο σε υψηλές συχνότητες, ενώ άλλα όπως η λύπη σε χαμηλές συχνότητες. Επίσης, συναισθήματα, όπως θυμός, χαρά και απαρέσκεια, παρουσιάζουν μεγάλες τιμές μέσης Teager ενέργειας, ενώ άλλα, όπως η λύπη και το ουδέτερο, έχουν μικρές τιμές.

- Το χαρακτηριστικό του *σταθμισμένου μέσου όρου στιγμιαίας συχνότητας* F παρουσιάζει διαφορές σε όλα τα συναισθήματα και σε όλα τα κανάλια συχνοτήτων από 0-6000Hz. Ιδιαίτερη διακριτότητα υπάρχει στις τιμές του F στο διάστημα 3500-5000Hz, όπως φαίνεται στο σχήμα 5.10α, αλλά και γύρω από το 1ο formant, όπως προκύπτει από το σχήμα 5.10β. Μεγάλη μεταβλητότητα του θυμού υπάρχει στις υψηλές συχνότητες, ενώ μεγάλη μεταβλητότητα της λύπης βρίσκεται στις χαμηλές συχνότητες.

- Παρόμοια με το F, η *σταθμισμένη απόκλιση της στιγμιαίας συχνότητας* B διαφοροποείται στα συναισθήματα σε όλες τις ζώνες συχνοτήτων. Παρόλα αυτά, όπως προκύπτει από την οπτική αναπαράσταση, δεν έχει τόσο μεγάλη διακριτότητα όση το χαρακτηριστικό B.

- Το *TEO-Auto-Env* είναι το πιο διακριτό από τα AM-FM χαρακτηριστικά διαμόρφωσης. Οι μεγαλύτερες διαφορές ανάμεσα στα συναισθήματα βρίσκονται στις σχυνότητες 0-2000Hz. Στις μεγαλύτερες συχνότητες υπάρχει μεμονωμένη πληροφορία, κυρίως για τη λύπη και το θυμό, όπως φαίνεται στα box plots του σχήματος 5.18.

- Η παραπάνω παρατήρηση δεν ισχύει για το *TEO-Pitch*, για το οποίο υπάρχει μεγαλύτερη διακριτότητα στις υψηλές συχνοτητες 4000-6500Hz, ενώ στις χαμηλές συχνότητες δε βλέπουμε ιδιαίτερες διαφορές μεταξύ των συναισθημάτων.

Κεφάλαιο 6

Ταξινόμηση των Συναισθημάτων με τη χρήση του Αλγορίθμου **K-means**

Ως μία πρώτη προσπάθεια ταξινόμησης των συναισθημάτων με βάση τα χαρακτηριστικά που περιγράψαμε στα προηγούμενα κεφάλαια, δοκιμάζουμε το μη επιβλεπόμενο αλγόριθμο K-means. Λόγω της απλότητας του αλγορίθμου εξετάζουμε την ταξινόμηση των συναισθημάτων ανά δύο και με τον τρόπο αυτό ελέγχουμε και τις πιθανές συσχετίσεις μεταξύ τους.

Διεξάγουμε τα πειράματα για τη βάση δεδομένων Berlin Database of Emotional Speech (Emo DB). Αρχικά, πειραματιζόμαστε μόνο με τον K-means χωρίς περαιτέρω επεξεργασία των χαρακτηριστικών. Στη συνέχεια, αυξάνουμε την πολυπλοκότητα ομαλοποιώντας τα AM-FM χαρακτηριστικά διαμόρφωσης, εφαρμόζοντας LDA πριν από τον K-means και τέλος, κάνοντας την κατάλληλη αρχικοποίηση για κάθε ομάδα.

Παρακάτω, δίνουμε μία σύντομη πεγραφή του K-means ενώ στη συνέχεια παραθέτουμε τα αποτελέσματα των πειραμάτων και τα συμπεράσματα που προκύπτουν από αυτά.

6.1 Σύντομη Περιγραφή του Αλγορίθμου **K-means**

Ο αλγόριθμος K-means [4] ομαδοποιεί ένα σύνολο δεδομένων $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ του ευκλείδιου χώρου διάστασης D σε K ομάδες. Κάθε ομάδα αντιπροσωπεύεται από ένα κέντρο μ_k, όπου $k = 1, \ldots K$. Επίσης, εισάγεται ένας δείκτης $r_{nk} \in \{0, 1\}$ που δείχνει σε ποια από τις K ομάδες ανήκει το σημείο $\mathbf{x}_n$, έτσι ώστε αν το $\mathbf{x}_n$ ανήκει στην ομάδα k ισχύει $r_{nk} = 1$, αλλιώς ισχύει $r_{nk} = 0$. Η συνάρτηση κριτηρίου που δείχνει την παραμόρφωση του συνόλου ορίζεται από τη σχέση

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2$$

Ο αλγόριθμος K-means βρίσκει τις τιμές των r_{nk} και μ_k, ώστε να ελαχιστοποιηθεί το κριτήριο J, ως εξής:

$$r_{nk} = \left\{ \begin{array}{c} 1, \quad \text{αν } k = arg\min_j \|\mathbf{x}_n - \mu_j\| \\ 0, \quad \text{αλλιώς} \end{array} \right\}$$

$$\mu_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}}$$

Οι δύο φάσεις ανάθεσης των σημείων $\mathbf{x}_n$ σε ομάδες και υπολογισμού των κέντρων μ_k εναλλάσσονται, έτσι ώστε να μην υπάρχει κάποια αλλαγή στις αναθέσεις (ή μέχρι να ξεπεραστεί ένα όριο επαναλήψεων).

Κάθε πρόταση που αντιπροσωπεύει ένα συναίσθημα χωρίζεται σε frames μήκους $25ms$ (400 δειγμάτων) με επικάλυψη $7.5ms$ (120 δείγματα). Για κάθε frame έχουν υπολογιστεί AM-FM, mfcc χαρακτηριστικά σε 4, 13, 20 και 30 κανάλια, pitch και formants. Το ποσοστό επιτυχίας p_k κάθε συναισθήματος ισούται με ποσοστό των frames όλων των προτάσεων που αντιστοιχούν στο συναίσθημα αυτό και έχουν ταξινομηθεί στη σωστή ομάδα, δηλαδή

$$p_k = \frac{\sum_{s=1}^{S_k} \sum_{f=1}^{F_{sk}} e_{sf}}{S_k \cdot F_{sk}}, \quad k = 1, \ldots, K$$

όπου S_k το πλήθος των προτάσεων του συναισθήματος k, F_{sk} το πλήθος των frames της πρότασης s για το συναίσθημα k, $e_{sf} = 1$ αν το frame f της πρότασης s έχει ταξινομηθεί στο συναίσθημα k, αλλιώς $e_{sf} = 0$. Στους επόμενους πίνακες απεικονίζουμε το μέσο όρο των ποσοστών p_k σωστής ταξινόμησης για όλα τα συναισθήματα $p = \frac{\sum_{k=1}^{K} p_k}{K}$.

Εκτός από το ποοσοστό επιτυχίας υπολογίζονται και οι αποστάσεις μεταξύ των κέντρων που δημιουργούνται από τον K-means, καθώς και οι αποστάσεις όλων των σημείων μίας ομάδας από το μέσο της. Έστω ότι θέλουμε να διαχωρίσουμε τα δεδομένα σε 2 κλάσεις με τη βοήθεια του K-means, οπότε $K = 2$. Ορίζουμε τη διαταξική απόσταση (inter distance) ως την ευκλείδια απόσταση μεταξύ των κέντρων των 2 κλάσεων:

$$d_{inter} = \|\mu_1 - \mu_2\|$$

Η ενδοταξική απόσταση (intra distance) ορίζεται ως η μέση τιμή των αποστάσεων των σημείων που ανήκουν στην ομάδα k από το μέσο της:

$$d_{intra,k} = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} \|x_n - \mu_k\|$$

όπου N_k είναι το πλήθος των σημείων που ανήκουν στην κλάση k. Ο λόγος ενδοταξικής δια διαταξική απόσταση (intra-inter distance rate) είναι το πηλίκο των δύο αυτών αποστάσεων. Αν έχει μικρή τιμή για μία ομάδα k, σημαίνει ότι η ομάδα αυτή είναι συμπαγής και απομονωμένη από τις υπόλοιπες ομάδες.

$$r_k = \frac{d_{intra,k}}{d_{inter}}$$

6.2 Πειραματικά Αποτελέσματα Ταξινόμησης με K-means

Στον πίνακα 6.1 φαίνονται τα αποτελέσματα της ταξινόμησης από τον αλγόριθμο K-means. Παρατηρούμε ότι στα περισσότερα χαρακτηριστικά όσο αυξάνεται ο αριθμός των καναλιών βρίσκουμε και καλύτερα ποσοστά επιτυχίας. Τα πιο αποδοτικά χαρακτηριστικά είναι ο μέσος όρος και η τυπική απόκλιση του στιγμιαίου πλάτους, ο μέσος όρος στιγμιαίας συχνότητας και το Auto-Env, που φαίνεται να είναι και το πιο ισχυρό. Είναι αξιοσημείωτο ότι στο TEO-Auto-Env η καλύτερη ταξινόμηση γίνεται για 13 κανάλια, αντί για 20 ή 30 που ισχύει σε άλλα χαρακτηριστικά, και αποφέρει παρόμοια αποτελέσματα. Αυτό πιθανώς σημαίνει ότι το TEO-Auto-Env είναι πιο περιεκτικό από τα υπόλοιπα χαρακτηριστικά.

6.2.1 Ταξινόμηση με K-means με κατάλληλη αρχικοποίηση

Αρχικοποιούμε τα κέντρα του αλγορίθμου K-means με το μέσο όρο των διανυσμάτων που ανήκουν σε κάθε ομάδα, έτσι ώστε να υπάρχει καλύτερη σύγκλιση του αλγορίθμου. Βλέπουμε ότι υπάρχουν ελαφρές βελτιώσεις των ποσοστών επιτυχίας, που είναι πιο εμφανείς στο FMP, το TEO-Auto-Env και το TEO-Pitch. Επίσης, όσο αυξάνεται το πλήθος των καναλιών, παρατηρούμε ότι η αρχικοποίηση αυτή επιφέρει σχεδόν παρόμοια αποτελέσματα με τον απλό K-means.

6.2.2 Ταξινόμηση με K-means σε Ομαλοποιημένα Χαρακτηριστικά

Εφαρμόζουμε ομαλοποίηση μετακινούμενου μέσου όρου με παράθυρο 50 δειγμάτων (μετά την εξάλειψη των NaNs που έχει ήδη γίνει) στα χαρακτηριστικά και δοκιμάζουμε να ταξινομήσουμε τα συναισθήματα με K-means. Τα αποτελέσματα φαίνονται στον πίνακα 6.1. Τα ομαλοποιημένα χαρακτηριστικά, ιδιαίτερα αυτά που προκύπτουν από τον αλγόριθμο E-SA, εμφανίζουν σε γενικές γραμμές καλύτερη απόδοση σε σχέση με τα μη ομαλοποιημένα. Επιτυγχάνονται μάλιστα καλύτερα αποτελέσματα με μικρότερο πλήθος καναλιών. Για παράδειγμα, παρατηρούμε ότι ο μέσος όρος και η τυπική απόκλιση του μη ομαλοποιημένου στιγμιαίου πλάτους είχαν καλύτερη απόδοση σε σύνολο 20 και 30 καναλιών. Τα αντίστοιχα ομαλοποιημένα χαρακτηριστικά ξεπερνούν τα προηγούμενα ποσοστά επιτυχίας με σύνολο μόνο 13 καναλιών. Με τον τρόπο αυτό επιβεβαιώνεται η υπόθεσή μας ότι η χρήσιμη πληροφορία του συναισθήματος στο λόγο εμπεριέχεται σε στοιχεία που 'σκιαγραφούν' το σήμα της φωνής χωρίς να αναδεικνύουν όλες τις λεπτομέρειες.

6.2.3 Ταξινόμηση με K-means σε Χαρακτηριστικά που έχουν υποστεί LDA

Πριν ταξινομήσουμε τα συναισθήματα με K-means εφαρμόζουμε LDA με στόχο τη μείωση των αλληλεπικαλύψεων μεταξύ των ομάδων και τα αποτελέσματα από το πείραμα αυτό φαίνονται στον πίνακα 6.1. Μεγάλη βελτίωση στην αναγνώριση έχει η μέθοδος LDA όταν εφαρμοστεί σε χαρακτηριστικά formants και TEO-Auto-Env. Στα υπόλοιπα χαρακτηριστικά δεν επιφέρει βελτίωση και μάλιστα χειροτερεύει πολλές φορές τα ποσοστά αναγνώρισης. Από αυτό μπορούμε να υποθέσουμε ότι τα χαρακτηριστικά formants και TEO-Auto-Env είναι πολύ διαχωρίσισμα με βάση τη μέθοδο αυτή, για αυτό και παρουσιάζουν επιτυχία στην αναγνώριση συναισθήματος. Είναι αξιοσημείωτο ότι το TEO-Auto-Env για 13 κανάλια έχει μέσο όρο ποσοστού επιτυχίας 93.84% στην ταξινόμηση των συναισθημάτων ανά δύο.

Κανάλια	4	13	20	30	4	13	20	30
Χαρακτηριστικό					Αρχικοποίηση			
formants	55.78				55.78			
mfcc	53.29	53.64	53.79	53.90	53.29	53.64	53.79	53.89
Energy Mean	55.71	55.82	57.57	57.98	57.58	57.49	57.54	57.69
Energy Std	55.50	55.61	57.56	57.65	57.44	57.39	57.58	57.78
Ampl Mean	55.50	57.35	59.97	60.97	57.33	59.80	59.97	60.76
Ampl Std	55.51	57.96	60.23	59.85	57.35	60.60	60.22	59.85
Freq Mean	52.93	53.30	57.09	57.83	53.90	54.41	57.30	57.83
Freq Std	52.36	53.82	57.27	53.28	53.14	55.09	57.86	53.27
B	53.17	54.15	53.98	54.78	54.22	55.53	53.98	54.78
F	52.70	53.42	53.11	52.27	53.59	54.57	53.51	52.10
FMP	55.03	53.92	57.75	56.92	56.70	55.23	57.75	56.92
TEO-Auto-Env	57.57	60.34	57.86	58.02	62.24	57.57	57.85	58.02
TEO-Pitch	59.25	55.25	57.35		63.86	62.18	55.68	

Κανάλια	4	13	20	30	4	13	20	30
Χαρακτηριστικό	Ομαλοποίηση				ΛΔΑ			
formants	59.48				59.16			
mfcc	54.17	55.13	55.78	56.68	53.88	72.16	74.69	75.61
Energy Mean	63.48	73.23	58.44	58.51	58.28	57.78	58.37	58.82
Energy Std	64.46	69.18	56.01	55.93	57.54	58.44	57.73	64.57
Amp Mean	67.11	76.20	65.94	67.73	59.95	65.74	66.77	58.82
Amp Std	68.02	76.45	67.34	67.96	62.12	66.32	67.30	58.95
Freq Mean	63.48	73.23	58.44	58.51	56.64	65.22	63.76	66.19
Freq Std	60.05	68.40	68.99	60.67	54.79	59.89	61.66	68.47
B	61.70	68.34	61.07	62.38	56.74	60.11	61.76	59.47
F	65.36	75.31	54.98	60.22	58.26	67.92	64.46	60.17
FMP	62.82	68.53	66.80	59.95	57.34	60.23	61.11	66.65
TEO-Auto-Env	63.16	60.34	58.41	58.24	87.98	93.84	72.05	60.14
TEO-Pitch	63.86	62.18	55.68		61.09	62.62	62.72	

Πίνακας 6.1: Μέσος όρος των ποσοστών επιτυχίας του αλγορίθμου K-means (K=2) για την ταξινόμηση συναισθημάτων ανά δύο α) με κατάλληλη αρχικοποίηση β)με ομαλοποίηση χαρακτηριστικών γ)με χρήση LDA.

6.2.4 Συνδυασμοί των Παραπάνω Τεχνικών

Επιχειρούμε ταξινόμηση με τον αλγόριθμο K-means συνδυάζοντας τις μεθόδους που αναφέραμε παραπάνω. Πιο συγκεκριμένα κάνουμε τους εξής συνδυασμούς, που φαίνονται και στον πίνακα 6.3: α) εφαρμογή του K-means με κατάλληλη αρχικοποίηση σε ομαλοποιημένα χαρακτηριστικά β)εφαρμογή του K-means σε ομαλοποιημένα χαρακτηριστικά που έχουν υποστεί LDA γ)εφαρμογή του K-means με κατάλληλη αρχικοποίηση σε χαρακτηριστικά που έχουν υποστεί LDA γ)εφαρμογή του K-means με κατάλληλη αρχικοποίηση σε ομαλοποιημένα χαρακτηριστικά που έχουν υποστεί LDA.

Παρατηρούμε ότι ο πιο δυνατός συνδυασμός είναι η ταξινόμηση των ομαλοποιημένων χαρακτηριστικών που έχουν υποστεί LDA, που επιφέρει πολύ καλύτερα αποτελέσματα απ'οτι αν εφαρμοστεί κάθε τεχνική ξεχωριστά. Η αρχικοποιηση των κέντρων του K-means με το μέσο όρο των χαρακτηριστικών δεν έχει πάντα καλύτερα αποτελέσματα σε σχέση με την τυχαία αρχικοποίηση. Για παράδειγμα, ο συνδυασμός και των τριών τεχνικών δεν έχει σχεδόν καμία βελτίωση σε σχέση με το συνδυασμό ομαλοποίησης και LDA χωρίς αρχικοποίηση.

Τέλος, για να έχουμε ένα μέτρο σύγκρισης με τα επόμενα κεφάλαια, εκτελούμε την ταξινόμηση με K-means και των 7 συναισθημάτων. Τα αποτελέσματα φαίνονται στον πίνακα 6.2, όπου πιο αποδοτικά είναι τα χαρακτηριστικά υπολογισμένα σε 20 κανάλια. Αυτό πιθανώς γιατί στα 20 κανάλια είναι πολλά και αντισταθμίζουν την αδυναμία του αλγορίθμου ταξινόμησης. Επιτυγχάνονται πάντως ποσοστά αναγνώρισης έως και 55.08% με κυρίαρχο τον απλό και σταθμισμένο μέσο όρο και τυπική απόκλιση στιγμιαίας συχνότητας και την τυπική απόκλιση του στιγμιαίου πλάτους.

Κανάλια	4	13	20	30	4	13	20	30
Χαρακτηριστικό	Ομαλοποίηση+Αρχικοποίηση				Ομαλοποίηση+LDA			
formants	59.49				72.44			
mfcc	54.17	55.13	55.78	58.94	65.03	85.24	87.31	88.10
Energy Mean	57.43	56.69	54.04	54.81	67.50	74.19	76.74	78.73
Ampl Mean	60.88	66.78	65.94	67.73	72.82	84.95	86.44	88.17
Freq Mean	58.02	57.10	58.47	58.51	67.97	80.98	79.94	82.58
Energy Std	57.18	56.49	56.16	57.42	69.28	76.35	77.10	79.22
Ampl Std	62.94	68.05	67.35	67.96	74.03	85.27	86.54	88.10
Freq Std	54.16	54.80	69.01	60.68	63.40	74.54	77.42	77.82
B	56.97	55.42	61.07	62.38	65.60	74.46	78.30	78.23
F	57.54	56.92	54.98	60.23	70.47	83.75	81.31	84.56
FMP	57.84	57.45	66.80	59.96	67.09	74.70	78.05	77.75
TEO-Auto-Env	62.60	60.34	58.41	58.06	89.14	94.80	83.49	81.17

Κανάλια	4	13	20	30	4	13	20	30
Χαρακτηριστικό	Αρχικοποίηση+LDA				Ομαλοποίηση+Αρχικοποίηση +ΛΔΑ			
formants	59.16				72.44			
mfcc	57.24	72.16	74.70	75.61	65.03	85.25	87.31	88.11
Energy Mean	59.10	59.02	57.73	59.01	68.48	74.19	77.17	79.10
Ampl Mean	61.26	66.95	67.07	67.85	72.82	84.95	86.44	88.17
Freq Mean	56.64	65.23	63.77	64.57	67.97	80.97	79.94	82.58
Energy Std	58.77	59.29	57.74	59.08	70.05	75.79	77.11	79.26
Ampl Std	61.87	67.07	67.30	67.50	74.03	85.27	86.54	88.10
Freq Std	54.79	59.90	61.67	59.47	63.40	74.54	77.42	77.82
B	56.74	60.11	61.76	60.17	65.61	74.46	78.30	78.23
F	58.26	67.92	64.46	66.65	70.48	83.75	81.31	84.56
FMP	57.34	60.23	61.11	60.14	67.09	74.70	78.05	77.75
TEO-Auto-Env	90.61	94.25	72.08	64.74	90.50	94.80	83.65	81.17

Πίνακας 6.2: Μέσος όρος των ποσοστών επιτυχίας του αλγορίθμου K-means (K=2) για την ταξινόμηση συναισθημάτων ανά δύο με βάση *συνδυασμούς τεχνικών*.

Κανάλια	4	6	12	20	4	6	12	20	4	6	12	20
Χαρακτηριστικό	Απλός K-means				Ομαλ+LDA				Ομαλ+LDA+ Αρχικ			
mfcc	25.73	25.8	24.13	29.82	20.74	22.44	27.3	47.56	20.92	21.79	26.58	47.45
Energy-mean	14.29	14.29	14.29	14.29	23.54	23.56	22.78	20.29	22.7	23.76	27.12	24.57
Ampl-mean	14.3	14.91	16.03	19.43	29.02	30.59	29.54	30.81	31.58	27.11	30.65	28.42
Freq-mean	23.48	22.18	22.15	18.14	25.11	25.41	38.71	55.08	25.39	27.66	38.92	54.83
Energy-std	14.29	14.29	14.29	14.29	27.32	23.9	20.09	19.97	21.56	20.2	19.92	19.75
Ampl-std	15.13	16.58	19.35	17.55	30.07	30.8	39.27	42.64	26.65	29.22	39.27	35.73
Freq-std	20.08	20.99	18.25	20.89	22.78	21.76	37.62	42.18	22.67	24.67	36.67	49.24
B	23.09	20.73	17.95	19.41	20.07	16.65	29.88	47.65	18.47	16.78	31.39	48.79
F	23	24.97	22.54	20.24	24.57	25.86	40.04	47.34	30.81	26.64	40.35	54.48
Auto-Env	14.29	14.29	14.99	14.29	19.85	22.35	19.08	22.15	25.49	22.35	18.97	22.15
formants	28.49				33.14				31.85			

Πίνακας 6.3: Μέσος όρος των ποσοστών επιτυχίας του αλγορίθμου K-means (K=7) για την ταξινόμηση των 7 συναισθημάτων με βάση *συνδυασμούς τεχνικών*.

6.3 Συμπεράσματα

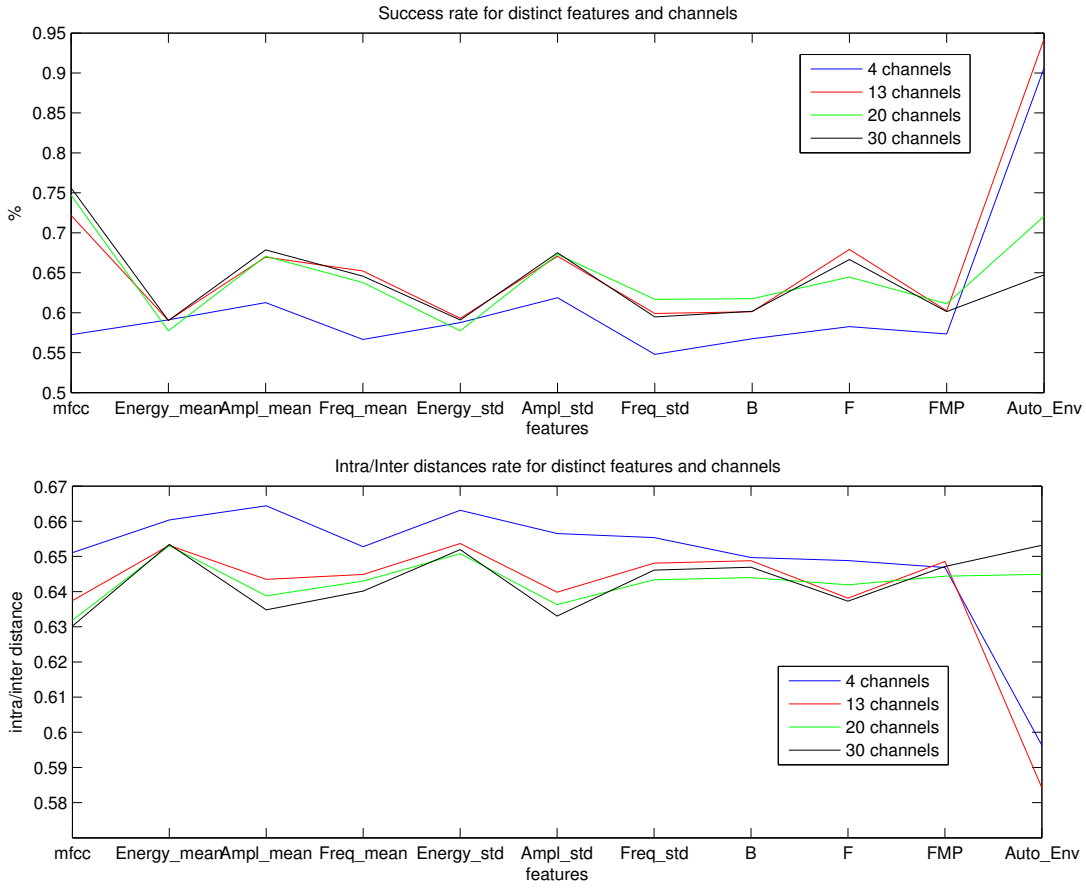
Εξάγουμε τα συμπεράσματα από την ταξινόμηση των συναισθημάτων με K-means κάνοντας 2 γραφικές παραστάσεις, που απεικονίζουν σε συγκεντρωτική μορφή τα αποτελέσματα. Στις παραστάσεις αυτές απεικονίζεται ο μέσος όρος τόσο του ποσοστού επιτυχίας όσο και του λόγου intra/inter distance των κλάσεων. Παρατηρούμε ότι τα δύο αυτά μεγάθη συμφωνούν, δηλαδή για μεγάλο ποσοστό επιτυχίας έχουμε μικρό λόγο intra/inter distance και αντίστροφα.

Στο σχήμα 6.1 απεικονίζεται το ποσοστό επιτυχίας και ο λόγος intra/inter distance για όλα τα κανάλια. Παρατηρούμε ότι τα χαρακτηριστικά υπολογισμένα σε 4 κανάλια, εξαιρουμένου του TEO-Auto-Env, δεν έχουν μεγάλη απόδοση. Αντίθετα, τα 13, 20 και 30 κανάλια εμφανίζουν παρόμοια ποσοστά επιτυχίας, πράγμα που υποδηλώνει ότι τα 13 κανάλια μπορεί να είναι πολλές φορές αρκετά για τη σωστή ταξινόμηση.

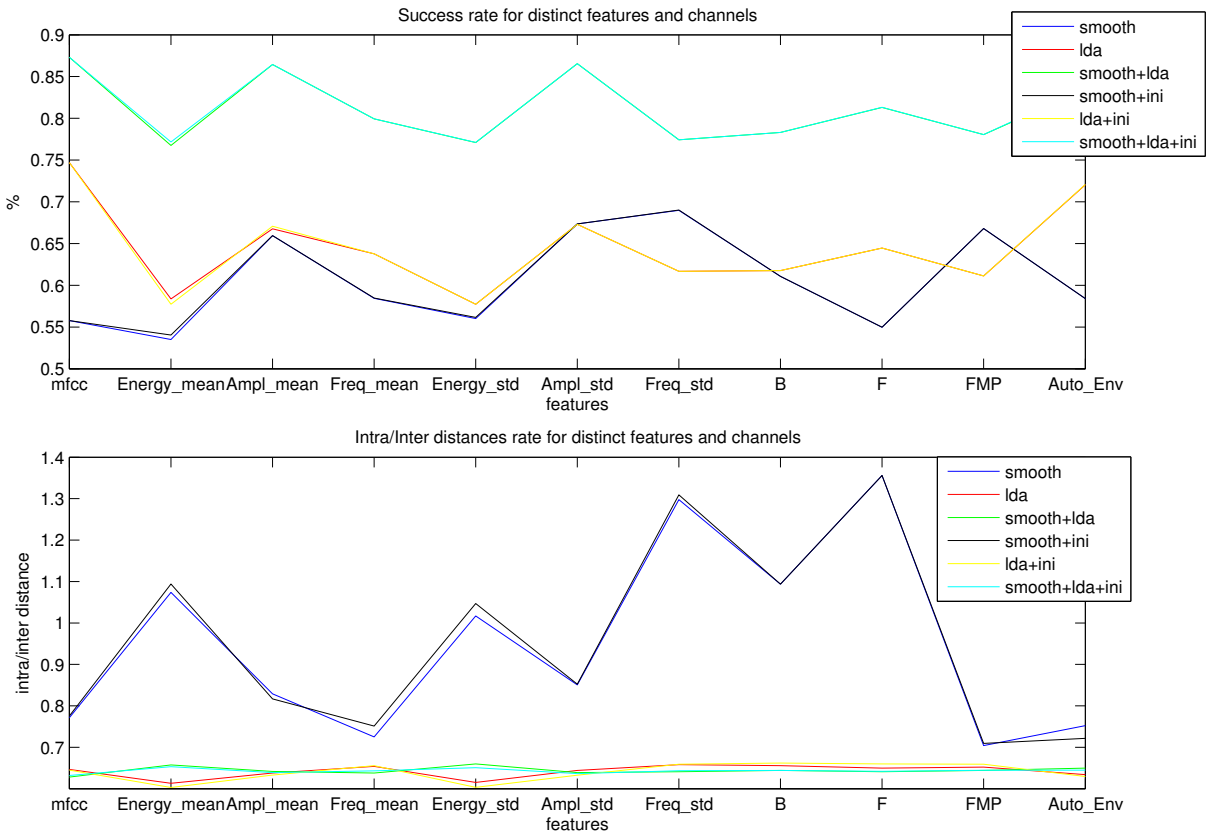
Επίσης, στο σχήμα 6.2 βλέπουμε τα ποσοστά επιτυχίας και το λόγο intra/inter distance για συνδυασμούς ομαλοποίησης, αρχικοποίησης και LDA. Τα ποσοστά δίνονται για όλα τα χαρακτηριστικά υπολογισμένα σε 20 κανάλια. Όπως φάνηκε και από τους αναλυτικούς πίνακες, ο συνδυασμός ομαλοποίησης και LDA είναι ο πιο αποδοτικός. Αυτό συμβαίνει πιθανώς γιατί η ομαλοποίηση αφαιρεί όλη την περιττή πληροφορία για την αναγνώριση συναισθήματος, ενώ η μέθοδος LDA είναι ειδικά σχεδιασμένη για τη μεγιστοποίηση της απόστασης μεταξύ των κλάσεων, που είναι βασικό κριτήριο για την καλή λειτουργία του K-means.

Όσον αφορά στη απόδοση των χαρακτηριστικών με βάση τον K-means, συμπεραίνουμε ότι το TEO-Auto-Env είναι το πιο ισχυρό χαρακτηριστικό. Ιδιαίτερη έμφαση πρέπει να δοθεί στο γεγονός ότι το χαρακτηριστικό αυτό παρουσιάζει καλύτερα αποτελέσματα σε 4 και 13 κανάλια και όχι σε 20 και 30. Σε συμφωνία με την παρατήρηση αυτή βρίσκονται και οι γραφικές παραστάσεις του προηγούμενου κεφαλαίου και ειδικά τα φασματογράμματα, που έδειχναν πολύ μικρές διαφορές μεταξύ των συναισθημάτων στις υψηλές συχνότητες για το χαρακτηριστικό αυτό. Ως προς την επιτυχία αναγνώρισης ακολουθούν τα χαρακτηριστικά του στιγμιαίου πλάτους, δηλαδή ο μέσος όρος και η τυπική απόκλιση σε κάθε frame. Αυτό δικαιολογείται από το γεγονός ότι το πλάτος και η ένταση του σήματος φωνής διαφοροποούνται ανάλογα με το συναίσθημα. Αρκετά αποδοτικά είναι και τα mfcc χαρακτηριστικά σε 30 κανάλια, που απεικονίζουν φασματικά χαρακτηριστικά, καθώς και τη δομή του φωνητικού σωλήνα. Τέλος, όπως αναμέναμε, επιτυχή αναγνώριση επιφέρουν και τα 4 πρώτα formants, αφού σχετίζονται με τη διαμόρφωση της πηγής του ηχητικού σήματος, που επηρεάζεται από το συναίσθημα.

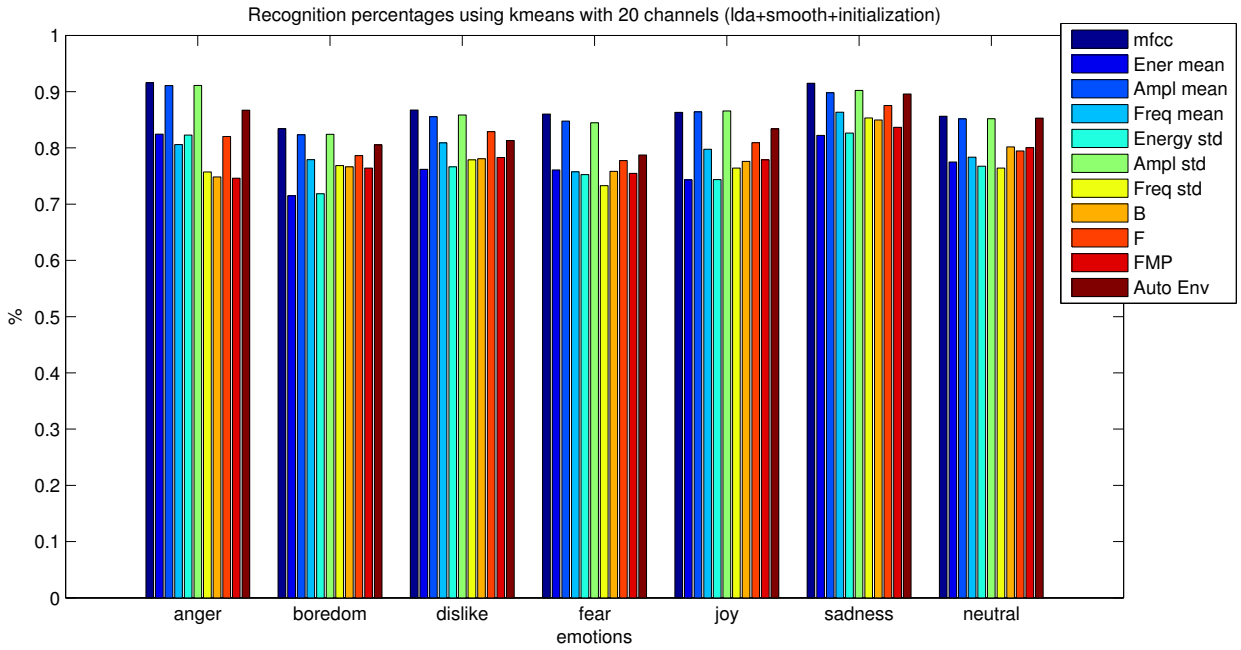
Τέλος, στο σχήμα 6.3 απεικονίζονται τα ποσοστά επιτυχίας για κάθε συναίσθημα σε όλα τα χαρακτηριστικά. Τα ποσοστά αυτά αντιστοιχούν στην ταξινόμηση με K-means με κατάλληλη αρχικοποίηση και με βάση ομαλοποιημένα χαρακτηριστικά που έχουν υποστεί LDA. Βλέπουμε ότι ο θυμός και η λύπη διακρίνονται καλύτερα από τα υπόλοιπα συναισθήματα. Μεσαία ποσοστά αναγνώρισης έχουν η απαρέσκεια, ο φόβος, η χαρά και το ουδέτερο, ενώ λιγότερο επιτυχής αναγνώριση γίνεται στην πλήξη.



Σχήμα 6.1: Ποσοστά επιτυχίας K-means (K=2) και λόγος intra/inter distance για όλα τα χαρακτηριστικά σε 4, 13, 20 και 30 κανάλια. Τα χαρακτηριστικά έχουν ομαλοποιηθεί και έχουν υποστεί LDA, ενώ στον K-means έχει εφαρμοστεί κατάλληλη αρχικοποίηση.



Σχήμα 6.2: Ποσοστά επιτυχίας K-means (K=2) και λόγος intra/inter distance για όλα τα χαρακτηριστικά σε 20 κανάλια για διάφορους συνδυασμούς τεχνικών.



Σχήμα 6.3: Ποσοστά επιτυχίας K-means (K=2) για όλα τα χαρακτηριστικά σε 20 κανάλια για όλα τα συναισθήματα.

Κεφάλαιο 7

Ταξινόμηση Συναισθημάτων με τη Χρήση Μείγματος Γκαουσιανών (GMMs)

7.1 Expectation maximization (EM) για μείγμα γκαουσιανών

Η μέθοδος ταξινόμησης με μείγματα γκαουσιανών στηρίζεται στην υπόθεση ότι ένα σύνολο δεδομένων μπορεί να αναπαρασταθεί από ένα μείγμα γκαουσιανών συναρτήσεων, οι παράμετροι του οποίου καθορίζονται μέσω Expectation Maximization (EM) [4, 26]. Σε κάθε κλάση συναισθήματος αντιστοιχεί μία γκαουσιανή. Εκτελούμε τα κλασσικά βήματα της μεθόδου expectation maximization ως εξής:

1. Αρχικοποίηση των παραμέτρων $\mathbf{m}_k$, $\mathbf{\Sigma}_k$ υπολογίζοντας το μέσο όρο και την τυπική απόκλιση των διανυσμάτων δεδομένων για κάθε κλάση-συναίσθημα. Ομοιόμορφη αρχικοποίηση του π_k για όλες τις κλάσεις.

2. **Ε-βήμα** Υπολογισμός του ποσοστού ανάθεσης του διανύσματος n στην κλάση k

$$w_{nk} = \frac{\pi_k \phi(\mathbf{x}_n|\mathbf{m}_n, \mathbf{\Sigma}_n)}{\sum_k \pi_k \phi(\mathbf{x}_n|\mathbf{m}_n, \mathbf{\Sigma}_n)}$$

   όπου

$$\phi(\mathbf{x}_n|\mathbf{m}_n, \mathbf{\Sigma}_n) = \exp[\frac{-||\mathbf{x_n} - \mathbf{m_k}||^2 - \min_k ||\mathbf{x_n} - \mathbf{m_k}||^2}{2\sigma_k^2}]$$

3. **Μ-βήμα** Επαναϋπολογισμός των παραμέτρων χρησιμοποιώντας τα ποσοστά ανάθεσης που βρέθηκαν στο Ε-βήμα

$$\mathbf{m}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \pi_{nk} \mathbf{x}_n$$

$$\mathbf{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} w_{nk} (\mathbf{x}_n - \mathbf{m}_k^{new})(\mathbf{x}_n - \mathbf{m}_k^{new})^T$$

$$\pi_k^{new} = \frac{N_k}{N}$$

όπου N είναι το πλήθος όλων των δεδομένων και

$$N_k = \sum_{n=1}^{N} \pi_{nk}$$

4. Υπολογισμός της πιθανοφάνειας

$$ln(p(\mathbf{X}|\mathbf{m}, \mathbf{\Sigma}, \pi)) = \sum_{n=1}^{N} ln \sum_{k=1}^{K} \pi_k \phi(\mathbf{x}_n|\mathbf{m}_n, \mathbf{\Sigma}_n) + \sum_{k=1}^{K} ln\pi_k N_k$$

όπου K το πλήθος των κλάσεων και X το σύνολο των παρατηρίσιμων δεδομένων. Αν η διαφορά της υπολογιζόμενης πιθανοφάνειας από αυτή που υπολογίστηκε στην προηγούμενη επανάληψη είναι μικρή(π.χ. 10^{-10}), τότε τερματίζει ο αλγόριθμος, αλλιώς συνεχίζει στο βήμα 2.

7.2 Μη Επιβλεπόμενη Ταξινόμηση με GMMs

Επιχειρούμε τη μη επιβλεπόμενη ταξινόμηση των 7 συναισθημάτων με τη χρήση GMMs 1 γκαουσιανής. Τα AM-FM χαρακτηριστικά διαμόρφωσης και τα MFCC έχουν υπολογιστεί σε 4, 6, 12 και 20 κανάλια.

Μία πρόταση που απεικονίζει ένα συναίσθημα θεωρείται ότι έχει ταξινομηθεί σωστά, αν η πλειοψηφία των παραθύρων της είναι σωστά ταξινομημένη (*majority success percentage*). Με βάση αυτόν τον υπολογισμό, παραθέτουμε τα ποσοστά επιτυχίας για κάθε χαρακτηριστικό σε 4, 6, 12 και 20 κανάλια. Στη συνέχεια, επιχειρούμε την ομαλοποίηση των χαρακτηριστικών, την εφαρμογή σε αυτά LDA και το συνδυασμό των μεθόδων αυτών. Τα αποτελέσματα για κάθε μέθοδο φαίνονται στους πίνακες 7.1 έως και 7.2.

Όπως παρατηρούμε, τα περισσότερα χαρακτηριστικά δίνουν μικρά ποσοστά επιτυχίας, τα οποία πιστεύεται ότι οφείλονται στη μη επιβλεπόμενη ταξινόμηση και όχι τόσο στα χαρακτηριστικά. Πάντως, σε συμφωνία με την ταξινόμηση με K-means, είναι φανερό ότι με εφαρμογή ομαλοποίησης και LDA στα χαρακτηριστικά ταξινόμησης, υπάρχει μεγάλη βελτίωση στην αναγνώριση. Επίσης, πιο αποδοτικός είναι ο υπολογισμός των χαρακτηριστικών σε 12 και 20 κανάλια.

Όσον αφορά στα χαρακτηριστικά, βλέπουμε ότι πολύ καλά ποσοστά επιτυχίας δίνουν τα χαρακτηριστικά συχνότητας, ειδικά όταν σε αυτά εφαρμοστεί ομαλοποίηση και LDA. Χωρίς να εφαρμοστεί ομαλοποίηση, αρκετά αποδοτικό είναι το εμβαδόν του στιγμιαίου πλάτους και τα formants. Η επίδραση της ομαλοποίησης είναι πιο εμφανής στα χαρακτηριστικά στιγμιαίας συχνότητας, γιατί πιθανώς είναι πιο θορυβώδη. Το εμβαδόν του στιγμιαίου πλάτους έχει έτσι κι αλλιώς αθροιστικό χαρακτήρα, που εξαλείφει τις απότομες μεταβολές, οπότε δεν προσδίδει κάτι παραπάνω η ομαλοποίηση στο χαρακτηριστικό αυτό.

Τέλος, δοκιμάζουμε αυτή τη μέθοδο ταξινόμησης με συνδυασμούς ανά 2 των χαρακτηριστικών και παραθέτουμε τα καλύτερα αποτελέσματα στον πίνακα 7.3. Επειδή, όπως είδαμε στην ταξινόμηση με βάση μεμονωμένα χαρακτηριστικά, μεγαλύτερη απόδοση έχει η εφαρμογή ομαλοποίησης και LDA, εφαρμόζουμε τις μεθόδους αυτές στα συνδυασμένα χαρακτηριστικά.

Παρατηρούμε ότι ορισμένοι συνδυασμοί αποφέρουν καλύτερα ποσοστά επιτυχίας, απ'οτι τα μεμονωμένα χαρακτηριστικά, καθώς και ότι ο υπολογισμός χαρακτηρισικών σε 12 κανάλια είναι ο πιο προσοδοφόρος για ταξινόμηση.

Ιδιαίτερη εντύπωση μας κάνει ο συνδυασμός της κύρτωσης του στιγμιαίου πλάτους (Ampl kurt) με τον απλό και σταθμισμένο μέσο όρο στιγμιαίας συχνότητας (Freq mean και F). Μεμονωμένα το Ampl kurt δίνει πολύ μέτρια αποτελέσματα, αλλά όταν συνδυαστεί με το Freq mean ή το F έχει πολύ καλή απόδοση, 62.82% και 62.77%. Το ποσοστό αυτό είναι καλύτερο και από το αυτό των μεμονωμένων Freq mean (55.05%) και F(53.96%). Όπως είχαμε παρατηρήσει και στα σχήματα σε προηγούμενο κεφάλαιο, η κύρτωση του στιγμιαίου πλάτους διαφέρει αρκετά ανάμεσα στα συναισθήματα, πράγμα που σημαίνει ότι τα συναισθήματα εμφανίζουν διαφορετικές μη κανονικές τιμές. Προφανώς, το μέγεθος της κύρτωσης από μόνο του δεν είναι ικανό να δώσει μεγάλη πληροφορία. Όμως, συνδυαζόμενο με ένα ισχυρό χαρακτηριστικό, όπως ο μέσος όρος στιγμιαίας συχνότητας, μπορεί να αποφέρει καλύτερα αποτελέσματα.

	απλό				**ομαλοποίηση**			
Χαρακτηριστικό	Αριθμός καναλιών				Αριθμός καναλιών			
	4	6	12	20	4	6	12	20
Energy-mean	14.29	14.29	14.29	14.29	14.29	14.29	13.4	21.23
Energy-std	14.29	14.29	14.29	14.29	14.29	14.29	14.17	18.11
Ampl-area	14.3	15.12	16.81	19.43	22.75	24.53	**29.1**	29.23
Ampl-kurt	14.29	14.29	14.29	14.29	26.77	28.11	17.36	16.64
Ampl-mean	14.29	14.64	15.9	17.95	21.01	23.71	**25.84**	24.88
Ampl-skew	24.42	28.92	14.29	21.53	22.55	23.9	20.64	23.19
Ampl-std	16.73	19.77	19.09	25.57	14.29	21.87	**29.25**	**29.87**
AmplDer1-mean	14.29	14.29	14.29	14.29	14.29	14.29	14.29	14.29
AmplDer1-std	14.29	14.29	14.29	14.29	14.29	14.29	14.29	14.29
AmplDer2-mean	14.29	14.29	14.29	14.29	14.29	14.29	14.29	14.29
AmplDer2-std	14.29	14.29	19.28	14.81	14.29	14.29	18.43	14.29
Auto-Env	14.29	14.29	14.29	14.29	18.1	19.72	20.44	20.68
B	23.37	20.44	17.59	20.59	19.05	18.62	18.59	25.28
F	22.55	24.77	22.27	18.79	25.4	24.67	**26.7**	24.18
FMP	18.74	17.23	19.48	14.06	16.72	17.95	17.48	20.54
Freq-mean	23.71	21.54	21.98	18.25	25.46	27.7	23.59	23.49
Freq-std	18.78	17.88	18.41	20.49	24.34	22.27	24.18	22.84
TEO-Pitch	24.93	16.4	14.29	14.29	**29.95**	22.39	19.57	22.05
mfcc	26.9	15.63	19.07	28.12	17.34	18.14	17	25.53
pitch	20.77				24.42			
formants	28.69				**30.49**			

Πίνακας 7.1: Μέσος όρος των ποσοστών επιτυχίας μη επιβλεπόμενης ταξινόμησης 7 συναισθημάτων με GMMs με βάση α) *απλά* χαρακτηριστικά και β)χαρακτηριστικά που έχουν υποστεί *ομαλοποίηση*.

	LDA				**ομαλοποίηση+LDA**			
Χαρακτηριστικό	Αριθμός καναλιών				Αριθμός καναλιών			
	4	6	12	20	4	6	12	20
Energy-mean	14.29	14.29	14.29	15.29	14.29	14.29	14.29	14.29
Energy-std	14.29	14.29	14.29	14.29	14.29	14.29	14.29	14.29
Ampl-area	14.99	16.7	19.42	21.23	31.26	27.33	31.04	28.07
Ampl-kurt	14.29	17.97	26.36	13.95	22.72	29.14	27.34	21.86
Ampl-mean	14.29	14.29	14.52	14.75	14.29	14.29	14.29	14.29
Ampl-skew	**29.29**	23.48	20.13	19.23	21.16	17.36	14.29	14.29
Ampl-std	14.29	14.29	14.29	14.29	14.29	14.29	14.29	14.29
AmplDer1-mean	14.29	14.29	14.29	14.29	14.29	14.29	14.29	14.29
AmplDer1-std	14.29	14.29	14.29	14.29	14.29	14.29	14.29	14.29
AmplDer2-mean	14.29	14.29	14.29	14.29	14.29	14.29	14.29	14.29
AmplDer2-std	14.29	14.29	14.29	14.29	14.29	14.29	14.29	14.29
Auto-Env	14.29	14.29	14.52	15.87	25.49	22.35	18.97	22.15
B	23.71	22.39	22.89	**31.11**	18.47	16.78	31.39	**48.79**
F	22.82	26.84	**37.06**	**34.32**	30.81	26.64	**40.35**	**53.96**
FMP	14.29	14.29	14.29	14.29	21.53	15.21	15.28	14.29
Freq-mean	25.42	21	**34.69**	31.38	25.39	27.66	39.1	**55.05**
Freq-std	22.86	21.32	25.3	27.3	22.67	24.67	36.67	**49.42**
TEO-Pitch	21.86	16.86	14.29	14.29	29.14	23.98	21.39	19.67
mfcc	16.36	27.48	14.52	40.54	14.29	14.29	14.29	14.29
pitch	20.77				24.42			
formants	25.43				**31.85**			

Πίνακας 7.2: Μέσος όρος των ποσοστών επιτυχίας μη επιβλεπόμενης ταξινόμησης 7 συναισθημάτων με GMMs με βάση χαρακτηριστικά που έχουν υποστεί α) *LDA* και β)*ομαλοποίηση+ LDA*.

Ομαλοποίηση+LDA							
Χαρακτηριστικό	Αριθμός καναλιών				Αριθμός καναλιών		
	4	6	12		4	6	12
Ampl-area+Ampl-kurt	26.64	32.91	**44.78**	Β+Φ	31.03	43.63	**61.02**
Ampl-area+Ampl-skew	29.61	40.5	35.36	Β+ΦΜΠ	14.29	15.21	15.59
Ampl-area+Auto-Env	34.68	31.62	43.8	Β+Φρεχ-μεαν	30.01	33.71	**59.18**
Ampl-area+B	32.93	31.51	38.85	Β+Φρεχ-στδ	37.61	29.86	35.99
Ampl-area+F	34.83	**44.16**	**51.76**	Β+μφςς	14.29	14.29	14.29
Ampl-area+Freq-mean	33.83	43.35	**54.23**	Β+πιτςη	24.08	18.24	15.29
Ampl-area+Freq-std	34.69	29.3	**44.36**	Β+φορμαντς	22.63	40.3	39.4
Ampl-area+TEO-Pitch	35.79	34.54	31.06	Φ+ΦΜΠ	14.29	15.24	15.28
Ampl-area+pitch	28.37	36.97	29.53	Φ+Φρεχ-μεαν	36.07	35.39	**54.86**
Ampl-area+formants	27.72	28.88	34.5	Φ+Φρεχ-στδ	38.44	48.29	**59.57**
Ampl-kurt+Auto-Env	36	43.17	**50.51**	Φ+ΤΕΟ-Πιτςη	40.74	45.19	34.15
Ampl-kurt+B	35.16	**44.64**	37.88	Φ+πιτςη	30.57	37.28	38.72
Ampl-kurt+F	43.32	26.09	**62.77**	Φ+φορμαντς	35.46	46.52	45.97
Ampl-kurt+Freq-mean	44.14	26.83	**62.82**	Φρεχ-μεαν+Φρεχ-στδ	34.47	47.79	**59.16**
Ampl-kurt+Freq-std	34.23	33	**45.43**	Φρεχ-μεαν+ΤΕΟ-Πιτςη	42.87	44.1	33.19
Ampl-kurt+TEO-Pitch	24.79	**44.58**	25.2	Φρεχ-μεαν+μφςς	25.19	14.29	14.29
Ampl-kurt+pitch	37.82	37.1	26.43	Φρεχ-μεαν+πιτςη	30.58	40.38	40.31
Ampl-kurt+formants	42.5	43.83	41.57	Φρεχ-μεαν+φορμαντς	33.62	29.41	41.82
Auto-Env+B	33.54	29.75	37.91	Φρεχ-στδ+ΤΕΟ-Πιτςη	22.43	34.04	32.65
Auto-Env+F	33.28	42.23	51.4	Φρεχ-στδ+πιτςη	25.85	34.48	**44.37**
Auto-Env+FMP	19.93	18.99	21.76	Φρεχ-στδ+φορμαντς	31.33	36.76	42.3
Auto-Env+Freq-mean	31.52	38.2	**50.94**	ΤΕΟ-Πιτςη+μφςς	25.08	14.29	14.29
Auto-Env+Freq-std	27.8	31.09	42.87	ΤΕΟ-Πιτςη+πιτςη	27.7	24.91	23
Auto-Env+TEO-Pitch	30.28	31.66	42.53	ΤΕΟ-Πιτςη+φορμαντς	32.2	28.58	30.59
Auto-Env+formants	28.46	32.28	36.27	πιτςη+φορμαντς	30.85	30.85	30.85

Πίνακας 7.3: Μέσος όρος των ποσοστών επιτυχίας μη επιβλεπόμενης ταξινόμησης με GMMs με βάση *συνδυασμούς χαρακτηριστικών* που έχουν υποστεί ομαλοποίηση και LDA.

7.3 Επιβλεπόμενη Ταξινόμηση με GMMs

Επιχειρούμε την επιβλεπόμενη ταξινόμηση των 7 συναισθημάτων με τη χρήση GMMs. Διεξάγουμε τα πειράματα στη βάση δεδομένων Berlin Database of Emotional Speech και χρησιμοποιούμε το 70% των προτάσεων κάθε συναισθήματος στο στάδιο της εκπαίδευσης και το 30% στο στάδιο της αναγνώρισης.

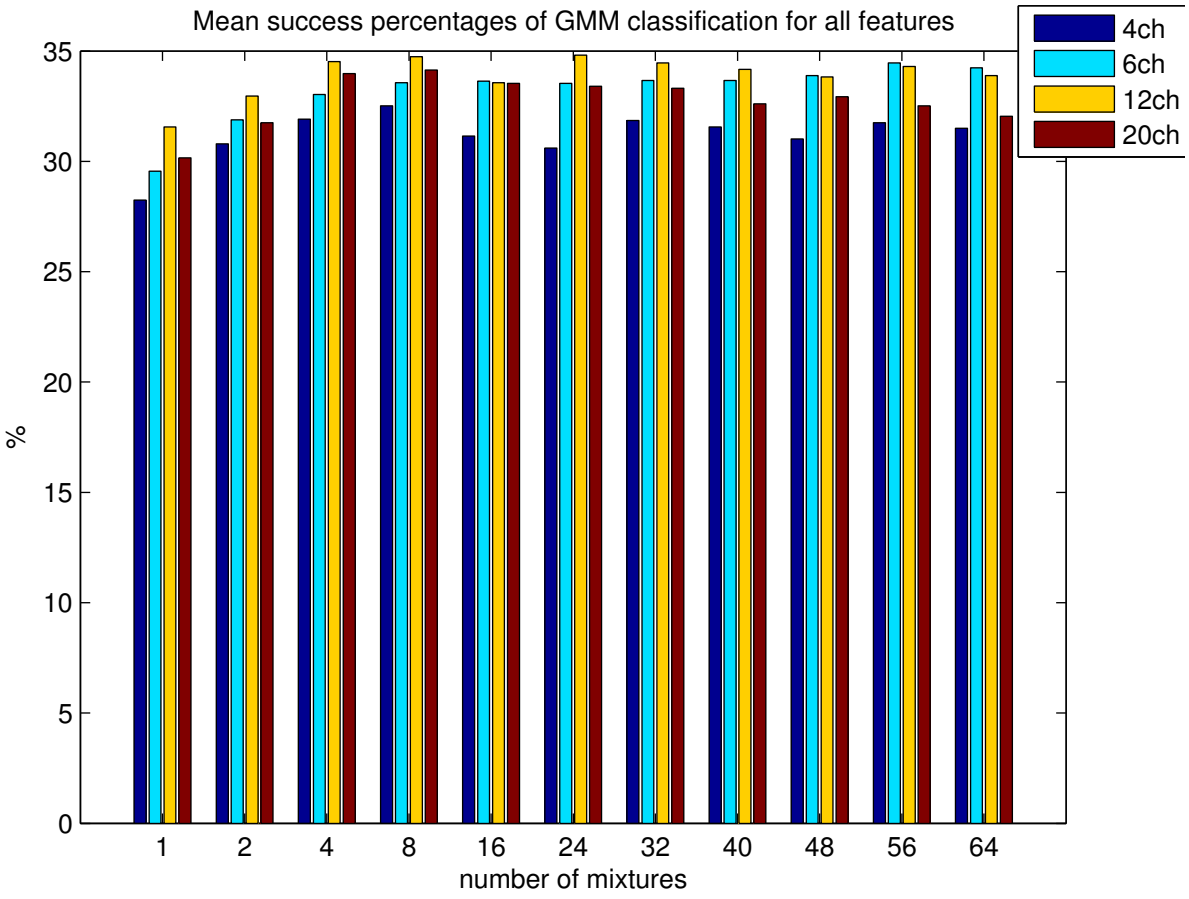
Αρχικά ταξινομούμε τα συναισθήματα με βάση μεμονωμένα χαρακτηριστικά υπολογισμένα σε 4, 6, 12 και 20 κανάλια. Για κάθε ομάδα-συναίσθημα δοκιμάζουμε πλήθος γκαουσιανών 1, 2, 4, 8, 16, 24, 32, 40, 48, 56 και 64. Τα αποτελέσματα φαίνονται στους πίνακες 7.4 έως 7.7.

Παρατηρούμε ότι με την αύξηση του πλήθους των γκαουσιανών, βελτιώνονται συνήθως τα αποτελέσματα αναγνώρισης. Επίσης, τις περισσότερες φορές, και κυρίως για τα AM-FM χαρακτηριστικά διαμόρφωσης, καλύτερα αποτελέσματα προκύπτουν για 12 κανάλια, παρά για 4, 6 ή ακόμα και 20.

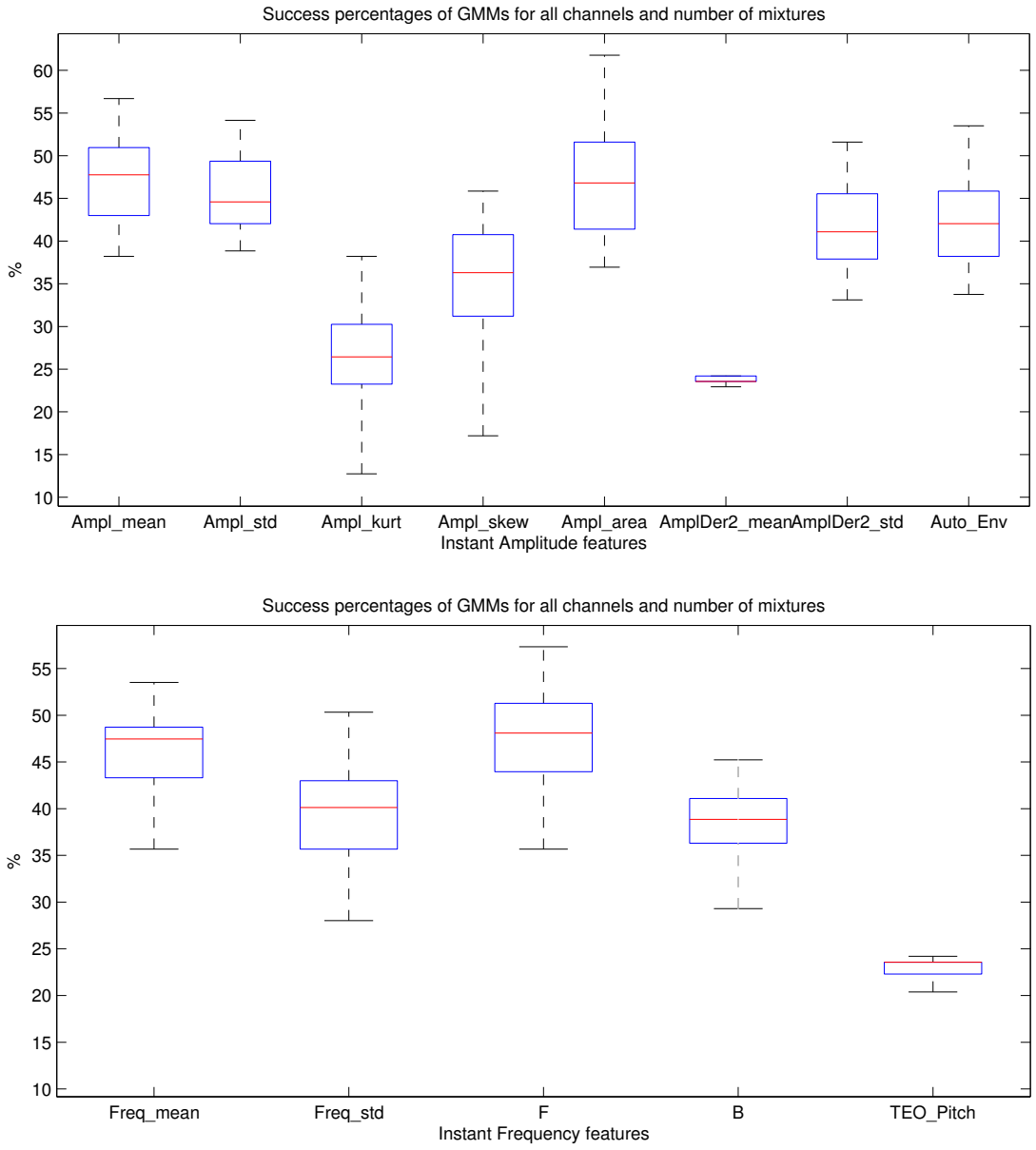
Όσον αφορά στην αποδοτικότητα των χαρακτηριστικών, βλέπουμε ότι το εμβαδόν, ο μέσος όρος και η τυπική απόκλιση του στιγμιαίου πλάτους (Ampl area, Ampl mean, Ampl std) εμφανίζουν μεγάλα ποσοστά επιτυχίας. Αυτό γιατί τα χαρακτηριστικά αυτά αναδεικνύουν τις μεταβολές της έντασης του σήματος, που είναι διακριτή ανάμεσα στα συναισθήματα. Επίσης, καλή απόδοση έχει ο απλός μέσος όρος και η σταθμισμένη απόκλιση της στιγμιαίας συχνότητας (Freq mean, F). Τα μεγέθη αυτά απεικονίζουν τις διαμορφώτριες συχνότητες, που καθορίζουν τη χρειά και τον τόνο της φωνής. Ιδιαίτερη εντύπωση μας κάνει το γεγονός ότι η τυπική απόκλιση της παραγώγου του στιγμαίου πλάτους που βασίζεται στη γκαουσιανή δίνει καλά αποτελέσματα. Πιθανώς αυτό συμβαίνει γιατί το χαρακτηριστικό αυτό αντικατοπτρίζει τις διακυμάνσεις του λόγου. Επίσης, λόγω του υπολογισμού της παραγώγου μέσω γκαουσιανής και όχι μέσω απλής διαφοράς χρονικών στιγμών, απεικονίζονται πιο ομαλά οι χρονικές μεταβολές, και εισάγεται με μεγαλύτερη ακρίβεια ο παράγοντας του χρόνου.

Στο σχήμα 7.1 απεικονίζονται οι μέσες τιμές των ποσοστών επιτυχίας σε όλα τα χαρακτηριστικά για κάθε κανάλι και κάθε αριθμό γκαουσιανών. Παρατηρούμε ότι τα χαρακτηριστικά υπολογισμένα σε 12 κανάλια είναι τα πιο αποδοτικά ανεξάρτητα του αριθμού των γκαουσιανών που χρησιμοοποιούνται στα GMMs. Επίσης, τα GMMs με 24, 8, 4 και 32 γκαουσιανές δίνουν καλύτερα αποτελέσματα από τα υπόλοιπα. Το πιθανότερο είναι ότι το πλήθος των 12 καναλιών είναι το ιδανικό, έτσι ώστε να διατηρείται η διακριτότητα μεταξύ των συχνοτήτων και παράλληλα να μη δίνεται έμφαση σε υπερβολικά μεγάλη λεπτομέρεια.

Στο σχήμα 7.2 βλέπουμε το εύρος των ποσοστών επιτυχίας για τα AM-FM χαρακτηριστικά διαμόρφωσης που προκύπτουν από το στιγμιαίο πλάτος και τη στιγμιαία συχνότητα. Από τα χαρακτηριστικά του στιγμιαίου πλάτους, ο μέσος όρος, η τυπική απόκλιση και το εμβαδόν του μεγέθους αυτού παρουσιάζουν μεγαλύτερη απόδοση, ενώ από τα χαρακτηριστικά της στιγμιαίας συχνότητας ο μέσος όρος και η σταθμισμένη τυπική απόκλιση έχουν πιο καλά ποσοστά επιτυχίας. Είναι άξιο προσοχής ότι το εμβαδόν του στιγμιαίου πλάτους (Ampl mean) έχει καλύτερη απόδοση από το εμβαδόν της αυτοσυσχέτισης του στιγμιαίου πλάτους (Auto Env). Αυτό σημαίνει ότι η πράξη της αυτοσυσχέτισης, που είχε προταθεί από τους Hansen, Patil [35], για την αναγνώριση άγχους, μπορεί να παραλειφθεί, οπότε να μειωθεί και η πολυπλοκότητα των υπολογισμών. Επίσης, οι μέσοι όροι στιγμιαίας συχνότητας (απλός και σταθμισμένος) είναι πιο αποδοτικοί από τις αποκλίσεις του μεγέθους αυτού. Με τον τρόπο αυτό, μπορεί να υποτεθεί ότι τα συναισθήματα εμφανίζουν παρόμοιες αποκλίσεις στιγμιαίας συχνότητας, αλλά όχι παρόμοιες τιμές.



Σχήμα 7.1: Μέση τιμή ποσοστών επιτυχίας GMMs με 1, 2, 4, 8, 16, 24, 32, 40, 48, 56 και 64 γκαουσιανές για όλα τα χαρακτηριστικά σε 4, 6, 12 και 20 κανάλια.



Σχήμα 7.2: Μέση τιμή ποσοστών επιτυχίας GMMs για όλα τα χαρακτηριστικά.

	1 γκαουσιανή				**2 γκαουσιανές**			
Χαρακτηριστικό	Αριθμός καναλιών				Αριθμός καναλιών			
	4	6	12	20	4	6	12	20
Energy mean	35.67	35.67	42.68	42.68	35.03	33.12	42.04	40.76
Energy std	39.49	40.13	42.04	41.40	39.49	40.13	43.95	43.31
Ampl area	44.59	41.40	48.41	49.68	42.04	41.40	50.96	47.77
Ampl kurt	12.74	11.47	12.10	12.74	22.93	24.20	31.21	14.65
Ampl mean	40.76	43.31	50.32	50.32	43.31	44.59	56.69	50.32
Ampl skew	19.11	19.11	17.20	15.29	42.68	45.86	26.11	25.48
Ampl std	38.85	41.40	38.85	40.123	46.50	47.77	39.49	43.31
AmplDer1 mean	24.20	24.20	24.20	24.20	24.20	24.20	24.20	24.20
AmplDer1 std	36.94	38.22	27.39	26.11	36.94	36.31	26.75	26.11
AmplDer2 mean	24.20	24.20	24.20	24.20	24.20	24.20	24.20	24.20
AmplDer2 std	33.12	40.76	43.95	43.31	36.31	40.13	45.86	45.22
TEO-Auto-Env	36.31	40.77	42.68	38.22	40.13	35.67	42.04	42.04
B	36.94	28.03	37.58	34.40	36.94	31.21	29.30	30.57
F	40.76	47.13	55.41	48.41	40.76	52.23	47.13	43.95
FMP	38.85	32.48	36.94	33.76	40.13	29.30	32.48	32.48
Freq mean	43.31	47.77	52.87	50.32	38.22	42.68	48.41	44.59
Freq std	19.11	28.03	34.40	24.20	34.40	41.40	43.95	45.22
TEO-Pitch	22.93	22.29	21.66	18.47	14.01	14.65	18.47	23.57
mfcc	17.20	17.20	15.29	19.11	17.83	17.83	18.47	19.75
M.O	**28.75**	**29.69**	**31.82**	**30.33**	**31.24**	**31.76**	**32.94**	**31.79**

Πίνακας 7.4: Μέσος όρος των ποσοστών επιτυχίας των GMMs με *1 και 2 γκαουσιανές ανά ομάδα* για την ταξινόμηση 7 συναισθημάτων.

	4 γκαουσιανές				**8 γκαουσιανές**				**16 γκαουσιανές**			
Χαρακτηριστικό	Αριθμός καναλιών				Αριθμός καναλιών				Αριθμός καναλιών			
	4	6	12	20	4	6	12	20	4	6	12	20
Energy mean	37.58	37.58	45.22	45.86	42.04	40.76	47.13	45.86	40.13	40.13	40.76	45.86
Energy std	41.4	39.49	43.95	47.13	41.4	41.4	49.68	45.86	36.31	36.94	40.13	39.49
Ampl area	38.85	42.04	47.77	45.22	36.94	51.59	45.86	42.68	38.85	48.41	54.78	53.5
Ampl kurt	25.48	27.39	21.02	26.11	24.2	22.93	27.39	23.57	24.84	30.57	28.03	31.85
Ampl mean	45.22	42.68	54.14	50.96	45.22	45.86	52.87	47.77	40.13	46.5	50.96	47.77
Ampl skew	31.21	41.4	31.21	32.48	36.31	43.95	33.76	26.75	36.31	39.49	28.66	28.03
Ampl std	53.5	49.68	42.04	43.31	54.14	52.87	44.59	43.95	52.23	49.04	42.04	41.4
AmplDer1 mean	24.2	24.2	24.2	24.2	24.2	24.2	24.2	24.2	23.57	23.57	23.57	23.57
AmplDer1 std	36.94	36.31	27.39	26.11	36.94	36.31	26.75	26.75	35.67	35.67	26.11	26.11
AmplDer2 mean	24.2	24.2	24.2	24.2	24.2	19.11	24.2	24.2	23.57	23.57	23.57	23.57
AmplDer2 std	35.03	41.4	47.13	45.22	35.03	42.04	51.59	45.86	33.76	40.13	49.04	40.13
TEO-Auto-Env	41.4	42.68	43.95	46.5	40.13	40.13	45.86	49.04	35.67	36.31	39.49	42.04
B	35.03	42.68	44.59	36.31	35.67	34.39	36.94	42.04	35.67	36.31	41.4	39.49
F	44.59	51.59	57.32	54.78	47.77	53.5	56.69	57.32	43.95	57.32	49.68	49.04
FMP	38.85	35.03	33.12	31.85	39.49	36.94	40.13	34.39				
Freq mean	47.77	44.59	53.5	52.87	49.04	47.13	53.5	50.32	47.77	46.5	45.86	47.77
Freq std	33.76	40.76	47.77	42.04	36.94	36.31	43.95	43.31	34.39	42.04	46.5	49.04
TEO-Pitch	24.2	14.01	20.38	17.2	22.29	21.66	11.46	24.2	23.57	23.57	23.57	23.57
mfcc	17.83	17.83	14.65	19.11	17.83	17.2	18.47	19.11	16.56	16.56	17.2	18.47
M.O	**32.24**	**33.12**	**34.46**	**33.88**	**32.85**	**33.72**	**35.00**	**34.15**	**31.15**	**33.63**	**33.57**	**33.54**

Πίνακας 7.5: Μέσος όρος των ποσοστών επιτυχίας των GMMs με *4 και 8 γκαουσιανές ανά ομάδα* για την ταξινόμηση 7 συναισθημάτων.

	24 γκαουσιανές				**32 γκαουσιανές**				**40 γκαουσιανές**			
Χαρακτηριστικό	Αριθμός καναλιών				Αριθμός καναλιών				Αριθμός καναλιών			
	4	6	12	20	4	6	12	20	4	6	12	20
Energy-mean	39.49	39.49	46.5	44.59	38.85	39.49	47.13	40.13	39.49	38.85	44.59	46.5
Energy-std	34.39	38.85	42.68	37.58	34.39	38.85	39.49	43.31	36.31	36.94	38.85	39.49
Ampl-area	36.94	45.22	59.24	52.87	40.76	45.86	61.78	53.5	38.22	43.95	59.24	51.59
Ampl-kurt	23.57	35.03	28.66	19.75	26.11	28.66	29.94	24.2	25.48	36.31	29.3	27.39
Ampl-mean	38.22	42.68	53.5	51.59	39.49	42.04	54.78	48.41	40.76	45.86	54.78	50.32
Ampl-skew	31.85	43.31	34.39	26.75	38.22	45.86	35.03	31.85	42.04	43.31	37.58	31.85
Ampl-std	50.32	47.77	40.76	41.4	50.96	47.13	41.4	41.4	51.59	44.59	42.68	43.31
AmplDer1-mean	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57
AmplDer1-std	35.03	35.67	26.11	26.11	35.67	35.67	26.11	26.11	35.03	35.67	26.11	26.11
AmplDer2-mean	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57
AmplDer2-std	33.76	40.13	48.41	43.31	33.76	39.49	49.04	42.04	33.76	39.49	47.77	42.68
Auto-Env	33.76	38.22	53.5	49.68	42.04	43.95	45.86	47.13	35.03	45.86	50.32	45.22
B	36.31	36.94	41.4	45.22	38.85	42.04	38.22	49.68	40.13	34.39	42.68	39.49
F	49.04	50.96	46.5	46.5	47.77	50.32	50.32	43.95	44.59	50.32	43.31	38.22
Freq-mean	47.77	48.41	46.5	43.31	47.13	47.77	43.95	43.95	48.41	47.77	40.76	37.58
Freq-std	34.39	41.4	40.76	50.32	35.67	38.85	39.49	42.04	33.12	42.68	39.49	43.95
TEO-Pitch	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57
mfcc	16.56	15.92	16.56	18.47	16.56	16.56	15.92	17.83	16.56	16.56	15.29	17.83
M.O	**30.61**	**33.54**	**34.81**	**33.41**	**31.85**	**33.66**	**34.46**	**33.31**	**31.56**	**33.66**	**34.17**	**32.61**

Πίνακας 7.6: Μέσος όρος των ποσοστών επιτυχίας των GMMs με *32 και 40 γκαουσιανές ανά ομάδα* για την ταξινόμηση 7 συναισθημάτων.

	48 γκαουσιανές				**56 γκαουσιανές**				**64 γκαουσιανές**			
Χαρακτηριστικό	Αριθμός καναλιών				Αριθμός καναλιών				Αριθμός καναλιών			
	4	6	12	20	4	6	12	20	4	6	12	20
Energy-mean	38.85	40.13	43.95	45.22	39.49	39.49	45.22	43.31	38.85	40.13	43.31	42.68
Energy-std	35.67	38.22	40.13	41.4	35.03	38.22	40.76	45.22	34.39	38.22	39.49	42.68
Ampl-area	37.58	43.95	56.05	51.59	38.22	40.76	56.69	50.32	38.85	49.04	56.69	50.32
Ampl-kurt	21.66	36.31	35.03	26.75	24.84	36.94	31.85	26.75	22.93	38.22	36.94	29.94
Ampl-mean	39.49	47.77	56.05	47.77	42.04	46.5	55.41	48.41	42.68	44.59	53.5	48.41
Ampl-skew	40.13	43.31	36.31	36.31	38.85	43.95	36.31	38.22	36.94	42.68	35.03	36.31
Ampl-std	50.96	47.13	44.59	43.31	52.23	47.77	45.22	45.22	50.96	47.13	42.68	43.31
AmplDer1-mean	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57
AmplDer1-std	35.03	35.67	26.11	26.11	35.03	35.67	26.11	26.11	35.03	35.67	26.11	26.11
AmplDer2-mean	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	22.93	23.57	23.57
AmplDer2-std	34.39	39.49	47.77	45.22	33.76	40.13	47.77	40.76	33.76	39.49	47.77	43.31
Auto-Env	33.76	36.31	40.13	49.68	35.67	40.76	49.68	47.77	36.31	38.85	49.68	43.95
B	37.58	40.13	38.85	43.31	40.76	40.76	42.04	38.85	37.58	38.85	40.13	38.85
F	43.95	50.96	43.95	35.67	49.68	55.41	42.68	36.94	50.96	54.78	41.4	35.67
Freq-mean	50.32	49.04	41.4	35.67	45.22	52.23	42.04	33.76	47.77	50.32	36.94	32.48
Freq-std	33.76	43.31	40.13	41.4	36.94	43.31	38.22	39.49	35.67	40.13	42.04	37.58
TEO-Pitch	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57	23.57
mfcc	16.56	15.29	15.29	18.47	16.56	16.56	15.29	18.47	16.56	16.56	15.29	18.47
M.O	**31.02**	**33.89**	**33.82**	**32.93**	**31.75**	**34.46**	**34.3**	**32.52**	**31.5**	**34.24**	**33.89**	**32.04**

Πίνακας 7.7: Μέσος όρος των ποσοστών επιτυχίας των GMMs με *48, 56 και 64 γκαουσιανές ανά ομάδα* για την ταξινόμηση 7 συναισθημάτων.

Στους πίνακες 7.9 έως και 7.13 φαίνονται τα αποτελέσματα της ταξινόμησης με GMMs 1 γκαουσιανής για συνδυασμούς χαρακτηριστικών. Παρατηρούμε ότι υπάρχουν συνδυασμοί που επιφέρουν καλύτερα αποτελέσματα σε σχέση με την ταξινόμηση με μεμονωμένα χαρακτηριστικά, όπως ο συνδυασμός F + Freq mean, που αποδίδει 61.78% επιτυχία. Επίσης, υπάρχουν συνδυασμοί μικρότερης διαστασιμότητας που έχουν ίδιο αποτέλεσμα, όπως το Freq mean + pitch (διάστασης 5), που δίνει ίδιο ποσοστό επιτυχίας (52.78%) με το Freq mean (διάστασης 12). Αυτό μπορεί να εξηγηθεί από το γεγονός ότι ο μέσος όρος στιγμιαίας συχνότητας σε 4 κανάλια αναπαριστά διαμορφώτριες συχνότητες και η πληροφορία που δίνει συμπληρώνεται από την πληροφορία της θεμελιώδους συχνότητας. Τέλος, μεγαλύτερα ποσοστά επιτυχίας απ'οτι τα μεμονωμένα χαρακτηριστικά σε 12 κανάλια, δίνουν οι συνδυασμοί B + F και B + Freq mean, όπου όλα τα χαρακτηριστικά υπολογίζονται σε 6 κανάλια. Τα ποσοστά επιτυχίας αυτών είναι 55.41% και 56.59% αντίστοιχα.

7.4 Συμπεράσματα

Από τα εκτελούμενα πειράματα, έχουμε δει σχετικά καλά ποσοστά επιτυχίας. Πιο καλά αποτελέσματα έχουν βρεθεί για 12 κανάλια και για ομαλοποιημένα χαρακτηριστικά που έχουν υποστεί LDA. Επίσης, ο πιο αποδοτικός ταξινομητής φαίνεται να είναι το μείγμα GMMs με 24 γκαουσιανές ανά ομάδα-συναίσθημα. Στον πίνακα 7.8 φαίνονται τα καλύτερα αποτελέσματα που έχουν προκύψει, τα χαρακτηριστικά αυτών και ο ταξινομητής από τον οποίο έχουν προκύψει. Τα αποτελέσματα είναι πολύ πιο υψηλά από αυτά που θα προέκυπταν από τυχαία επιλογή (100/7=14.3%). Από τα χαρακτηριστικά πλάτους πιο αποδοτικό είναι το Ampl area, ενώ από τα χαρακτηριστικά συχνότητας πιο ισχυρά φαίνεται ο μέσος όρος (σταθμισμένος ή μη). Ιδιαίτερη εντύπωση μας κάνει ότι ο συνδυασμός χαρακτηριστικών, αν ταξινομηθεί με GMM 1 γκαουσιανής, μπορεί να αποφέρει ίδιο αποτέλεσμα με μεμονωμένα χαρακτηριστικά που ταξινομούνται με GMM περισσότερων γκαουσιανών.

Χαρακτηριστικό	**Ταξινομητής**	**Ποσοστό %)**
Ampl-area12	ΓΜΜ32	61.78
Ampl-area12	ΓΜΜ40	59.24
F + Freq-mean12	ΓΜΜ1	61.78
F6	ΓΜΜ56	55.41

Πίνακας 7.8: Τα καλύτερα αποτελέσματα από την επιβλεπόμενη ταξινόμηση με GMMs.

	1 γκαουσιανή			
Χαρακτηριστικό	Αριθμός καναλιών			
	4	6	12	20
Energy-mean+Energy-std	38.85	36.94	40.13	42.68
Energy-mean+Ampl-area	43.95	42.04	45.22	40.76
Energy-mean+Ampl-kurt	26.75	29.94	36.94	41.4
Energy-mean+Ampl-mean	42.04	41.4	42.68	42.68
Energy-mean+Ampl-skew	34.39	37.58	41.4	40.76
Energy-mean+Ampl-std	40.13	42.68	47.13	42.04
Energy-mean+AmplDer1-mean	34.39	34.39	41.4	42.68
Energy-mean+AmplDer1-std	35.67	36.31	42.68	42.04
Energy-mean+AmplDer2-mean	34.39	34.39	42.04	41.4
Energy-mean+AmplDer2-std	36.94	37.58	45.22	41.4
Energy-mean+Auto-Env	35.03	40.76	41.4	33.12
Energy-mean+B	37.58	39.49	42.68	41.4
Energy-mean+F	35.03	40.13	48.41	42.68
Energy-mean+Freq-mean	38.22	39.49	48.41	39.49
Energy-mean+Freq-std	33.12	38.85	33.76	23.57
Energy-mean+TEO-Pitch	28.66	32.48	30.57	38.22
Energy-mean+mfcc	17.2	17.2	18.47	17.2
Energy-mean+pitch	37.58	38.85	42.68	41.4
Energy-mean+formants	36.31	36.94	41.4	42.04
Energy-std+Ampl-area	40.76	38.85	46.5	40.13
Energy-std+Ampl-kurt	19.75	24.2	31.85	35.67
Energy-std+Ampl-mean	42.04	40.76	48.41	40.76
Energy-std+Ampl-skew	31.21	29.3	34.39	35.67
Energy-std+Ampl-std	40.76	42.68	43.95	36.31
Energy-std+AmplDer1-mean	31.21	31.85	37.58	36.94
Energy-std+AmplDer1-std	36.94	38.85	39.49	35.67
Energy-std+AmplDer2-mean	31.21	31.85	38.22	35.67
Energy-std+AmplDer2-std	35.67	37.58	45.22	38.22
Energy-std+Auto-Env	35.67	40.13	39.49	33.12
Energy-std+B	34.39	39.49	41.4	40.76
Energy-std+F	37.58	40.76	46.5	41.4
Energy-std+Freq-mean	37.58	42.04	47.77	38.85
Energy-std+Freq-std	30.57	38.85	35.67	23.57
Energy-std+TEO-Pitch	24.2	23.57	23.57	29.3
Energy-std+mfcc	15.92	14.01	17.83	18.47
Energy-std+pitch	35.03	42.68	40.76	36.94
Energy-std+formants	40.13	35.67	39.49	35.03
Ampl-area+Ampl-kurt	40.13	37.58	41.4	43.31
Ampl-area+Ampl-mean	42.04	39.49	45.86	48.41
Ampl-area+Ampl-skew	42.68	42.04	47.13	45.22

Πίνακας 7.9: Μέσος όρος των ποσοστών επιτυχίας των GMMs με 1 γκαουσιανή ανά ομάδα και με βάση *συνδυασμούς χαρακτηριστικών* για την ταξινόμηση 7 συναισθημάτων.

	1 γκαουσιανή συνέχεια)			
Χαρακτηριστικό	Αριθμός καναλιών			
	4	6	12	20
Ampl-area+Ampl-std	42.68	38.22	46.5	45.22
Ampl-area+AmplDer1-mean	42.04	39.49	45.86	48.41
Ampl-area+AmplDer1-std	42.04	40.76	48.41	47.77
Ampl-area+AmplDer2-mean	42.04	39.49	45.86	47.77
Ampl-area+AmplDer2-std	42.04	40.13	46.5	49.04
Ampl-area+Auto-Env	34.39	39.49	37.58	29.3
Ampl-area+B	42.68	40.13	47.13	48.41
Ampl-area+F	45.22	44.59	52.23	48.41
Ampl-area+Freq-mean	43.31	44.59	**51.59**	38.22
Ampl-area+Freq-std	40.13	42.68	38.85	23.57
Ampl-area+TEO-Pitch	35.03	28.66	29.3	40.13
Ampl-area+mfcc	16.56	18.47	18.47	17.83
Ampl-area+pitch	41.4	43.95	48.41	47.77
Ampl-area+formants	41.4	40.76	47.77	47.77
Ampl-kurt+Ampl-mean	35.03	36.31	43.95	43.95
Ampl-kurt+Ampl-skew	15.92	16.56	18.47	20.38
Ampl-kurt+Ampl-std	24.2	33.76	38.22	37.58
Ampl-kurt+AmplDer1-mean	12.1	10.83	11.46	11.46
Ampl-kurt+AmplDer1-std	11.46	10.83	14.01	12.74
Ampl-kurt+AmplDer2-mean	12.1	10.83	11.46	11.46
Ampl-kurt+AmplDer2-std	12.1	12.74	38.85	38.22
Ampl-kurt+Auto-Env	32.48	37.58	38.22	31.21
Ampl-kurt+B	12.74	16.56	29.3	23.57
Ampl-kurt+F	19.11	28.03	37.58	36.31
Ampl-kurt+Freq-mean	22.93	27.39	37.58	34.39
Ampl-kurt+Freq-std	10.83	16.56	26.75	23.57
Ampl-kurt+TEO-Pitch	14.65	18.47	19.75	9.55
Ampl-kurt+mfcc	10.83	10.83	10.83	18.47
Ampl-kurt+pitch	16.56	18.47	19.11	20.38
Ampl-kurt+formants	15.92	17.83	14.65	14.01
Ampl-mean+Ampl-skew	41.4	40.76	45.86	46.5
Ampl-mean+Ampl-std	42.04	38.85	46.5	47.77
Ampl-mean+AmplDer1-mean	39.49	42.04	47.77	47.77
Ampl-mean+AmplDer1-std	41.4	38.85	47.13	47.13
Ampl-mean+AmplDer2-mean	39.49	42.04	47.77	47.77
Ampl-mean+AmplDer2-std	40.76	40.76	46.5	47.13
Ampl-mean+Auto-Env	39.49	40.13	43.95	32.48
Ampl-mean+B	42.68	43.31	**50.96**	49.04
Ampl-mean+F	44.59	44.59	**53.5**	49.04
Ampl-mean+Freq-mean	44.59	45.86	**51.59**	48.41

Πίνακας 7.10: Μέσος όρος των ποσοστών επιτυχίας των GMMs με 1 γκαουσιανή ανά ομάδα και με βάση *συνδυασμούς χαρακτηριστικών* για την ταξινόμηση 7 συναισθημάτων (συνέχεια).

	1 γκαουσιανή συνέχεια)			
Χαρακτηριστικό	Αριθμός καναλιών			
	4	6	12	20
Ampl-mean+Freq-std	38.22	44.59	45.86	23.57
Ampl-mean+TEO-Pitch	25.48	28.03	27.39	38.85
Ampl-mean+mfcc	17.2	18.47	18.47	18.47
Ampl-mean+pitch	42.04	45.86	48.41	48.41
Ampl-mean+formants	41.4	40.13	47.77	49.04
Ampl-skew+Ampl-std	40.13	41.4	39.49	40.76
Ampl-skew+AmplDer1-mean	17.83	17.83	15.92	14.01
Ampl-skew+AmplDer1-std	34.39	38.22	35.67	17.2
Ampl-skew+AmplDer2-mean	17.83	17.83	15.92	14.01
Ampl-skew+AmplDer2-std	35.03	40.13	41.4	42.68
Ampl-skew+Auto-Env	35.03	38.22	40.76	31.85
Ampl-skew+B	33.76	34.39	35.03	35.03
Ampl-skew+F	37.58	42.04	**50.96**	46.5
Ampl-skew+Freq-mean	41.4	47.77	48.41	45.22
Ampl-skew+Freq-std	15.92	32.48	32.48	23.57
Ampl-skew+TEO-Pitch	18.47	21.66	22.29	15.92
Ampl-skew+mfcc	17.2	17.83	17.83	18.47
Ampl-skew+pitch	30.57	29.94	40.13	33.76
Ampl-skew+formants	32.48	31.85	32.48	23.57
Ampl-skew+AmplDer1-mean	37.58	39.49	37.58	38.85
Ampl-skew+AmplDer1-std	36.31	43.31	37.58	38.85
Ampl-skew+AmplDer2-mean	37.58	39.49	37.58	38.85
Ampl-skew+AmplDer2-std	38.22	43.95	46.5	40.13
Ampl-skew+Auto-Env	39.49	38.85	40.76	36.94
Ampl-skew+B	42.68	46.5	42.04	40.76
Ampl-skew+F	48.41	48.41	47.13	42.68
Ampl-skew+Freq-mean	48.41	51.59	45.22	41.4
Ampl-skew+Freq-std	39.49	45.86	35.67	23.57
Ampl-skew+TEO-Pitch	22.29	27.39	22.29	18.47
Ampl-skew+mfcc	18.47	17.83	17.83	18.47
Ampl-skew+pitch	40.76	42.68	37.58	40.13
Ampl-skew+formants	43.31	43.31	37.58	39.49
AmplDer1-mean+AmplDer1-std	35.67	36.94	26.75	24.84
AmplDer1-mean+AmplDer2-mean	23.57	23.57	23.57	23.57
AmplDer1-mean+AmplDer2-std	31.21	39.49	42.04	42.04
AmplDer1-mean+Auto-Env	34.39	38.22	40.76	38.85
AmplDer1-mean+B	35.67	26.75	35.03	32.48
AmplDer1-mean+F	38.22	45.22	52.87	47.13
AmplDer1-mean+Freq-mean	40.13	45.22	49.68	49.68
AmplDer1-mean+Freq-std	18.47	27.39	38.22	36.31

Πίνακας 7.11: Μέσος όρος των ποσοστών επιτυχίας των GMMs με 1 γκαουσιανή ανά ομάδα και με βάση *συνδυασμούς χαρακτηριστικών* για την ταξινόμηση 7 συναισθημάτων (συνέχεια).

	1 γκαουσιανή συνέχεια)			
Χαρακτηριστικό	Αριθμός καναλιών			
	4	6	12	20
AmplDer1-mean+TEO-Pitch	22.29	21.02	21.02	17.83
AmplDer1-mean+mfcc	15.92	15.92	13.38	17.83
AmplDer1-mean+pitch	24.84	25.48	25.48	27.39
AmplDer1-mean+formants	30.57	30.57	31.21	31.21
AmplDer1-std+AmplDer2-mean	35.67	36.94	26.75	24.84
AmplDer1-std+AmplDer2-std	33.76	37.58	42.68	41.4
AmplDer1-std+Auto-Env	36.94	37.58	41.4	37.58
AmplDer1-std+B	40.76	41.4	33.76	28.03
AmplDer1-std+F	42.68	46.5	47.77	46.5
AmplDer1-std+Freq-mean	44.59	49.04	47.13	48.41
AmplDer1-std+Freq-std	22.93	38.85	38.85	12.1
AmplDer1-std+TEO-Pitch	21.02	22.29	21.02	17.83
AmplDer1-std+mfcc	17.83	18.47	10.83	17.83
AmplDer1-std+pitch	34.39	39.49	27.39	28.66
AmplDer1-std+formants	36.31	39.49	28.66	29.94
AmplDer2-mean+AmplDer2-std	31.85	39.49	42.04	42.04
AmplDer2-mean+Auto-Env	34.39	38.22	41.4	36.94
AmplDer2-mean+B	35.67	26.75	35.67	32.48
AmplDer2-mean+F	38.22	45.22	**52.87**	47.13
AmplDer2-mean+Freq-mean	40.13	45.22	49.68	49.68
AmplDer2-mean+Freq-std	18.47	27.39	33.76	23.57
AmplDer2-mean+TEO-Pitch	22.29	21.02	22.29	17.83
AmplDer2-mean+mfcc	15.92	15.92	14.01	17.83
AmplDer2-mean+pitch	24.84	24.84	24.84	24.84
AmplDer2-mean+formants	30.57	30.57	30.57	30.57
AmplDer2-std+Auto-Env	36.94	37.58	42.04	37.58
AmplDer2-std+B	42.68	46.5	44.59	44.59
AmplDer2-std+F	45.86	47.13	**53.5**	44.59
AmplDer2-std+Freq-mean	48.41	**50.32**	**50.96**	45.86
AmplDer2-std+Freq-std	29.94	44.59	42.68	23.57
AmplDer2-std+TEO-Pitch	18.47	21.02	22.93	19.11
AmplDer2-std+mfcc	17.2	18.47	18.47	18.47
AmplDer2-std+pitch	37.58	40.13	44.59	41.4
AmplDer2-std+formants	37.58	40.13	44.59	42.68
Auto-Env+B	35.03	38.22	40.76	29.94
Auto-Env+F	38.85	39.49	43.95	36.31
Auto-Env+Freq-mean	37.58	39.49	45.86	28.66
Auto-Env+Freq-std	36.31	36.94	31.85	23.57
Auto-Env+TEO-Pitch	35.67	38.22	34.39	33.12
Auto-Env+mfcc	17.83	18.47	16.56	11.46

Πίνακας 7.12: Μέσος όρος των ποσοστών επιτυχίας των GMMs με 1 γκαουσιανή ανά ομάδα και με βάση *συνδυασμούς χαρακτηριστικών* για την ταξινόμηση 7 συναισθημάτων (συνέχεια).

	1 γκαουσιανή συνέχεια)			
Χαρακτηριστικό	Αριθμός καναλιών			
	4	6	12	20
Auto-Env+pitch	37.58	39.49	42.04	36.31
Auto-Env+formants	36.94	38.85	42.04	34.39
B+F	45.22	**55.41**	**52.23**	40.76
B+Freq-mean	47.77	**56.69**	**53.5**	17.2
B+Freq-std	26.75	28.03	33.12	23.57
B+TEO-Pitch	21.02	26.11	21.02	19.11
B+mfcc	16.56	15.92	17.83	17.2
B+pitch	40.13	42.04	36.94	31.21
B+formants	38.85	36.31	35.03	32.48
F+Freq-mean	43.31	49.68	**61.78**	48.41
F+Freq-std	36.31	43.31	**54.78**	23.57
F+TEO-Pitch	22.29	29.3	27.39	24.84
F+mfcc	17.83	17.83	14.65	14.65
F+pitch	39.49	47.77	**52.87**	47.77
F+formants	45.22	45.22	46.5	45.22
Freq-mean+Freq-std	42.04	44.59	33.76	23.57
Freq-mean+TEO-Pitch	25.48	30.57	29.94	13.38
Freq-mean+mfcc	18.47	18.47	15.92	7.64
Freq-mean+pitch	41.4	**52.87**	**51.59**	49.04
Freq-mean+formants	41.4	45.22	49.68	47.77
Freq-std+TEO-Pitch	18.47	24.84	20.38	23.57
Freq-std+mfcc	17.83	17.83	7.64	7.64
Freq-std+pitch	35.03	38.22	35.67	23.57
Freq-std+formants	31.85	33.76	35.03	23.57
TEO-Pitch+mfcc	12.1	15.29	14.01	14.65
TEO-Pitch+pitch	26.11	25.48	22.29	17.83
TEO-Pitch+formants	24.2	24.2	22.93	17.83
mfcc+pitch	12.74	12.74	12.74	17.83
mfcc+formants	15.92	15.92	15.92	17.83

Πίνακας 7.13: Μέσος όρος των ποσοστών επιτυχίας των GMMs με 1 γκαουσιανή ανά ομάδα και με βάση *συνδυασμούς χαρακτηριστικών* για την ταξινόμηση 7 συναισθημάτων (συνέχεια).

Κεφάλαιο 8

Ταξινόμηση Συναισθημάτων με Αναπροσαρμοζόμενο Μείγμα Γκαουσιανών και Απόσταση Mahalanobis

8.1 Περιγραφή Αλγορίθμου

Ο αλγόριθμος ΕΜ για την εύρεση των παραμέτρων των γκαουσιανών χρησιμοποιείται ευρέως στη βιβλιογραφία. Ωστόσο υποφέρει από δύο βασικά μειονεκτήματα: α)ο αριθμός των γκαουσιανών είναι καθορισμένος από την αρχή β)η αρχικοποίηση των παραμέτρων επηρεάζει το τελικό αποτέλεσμα. Οι Ververidis, Kotropoulos πρότειναν μία εναλλακτική υλοποίηση του ΕΜ για μείγμα γκαουσιανών που χρησιμοποιεί την απόσταση Mahalanobis [92]. Η γενική ιδέα του αλγορίθμου είναι ότι αρχικά θεωρείται ότι το σύνολο των δεδομένων αποτελεί μία κλάση, διασπάται η κλάση αυτή σε δύο κλάσεις, στη συνέχεια σε τρεις, κοκ, ώσπου κάθε κλάση να μπορεί να προκύψει από μία μόνο γκαουσιανή.
Έστω το σύνολο δεδομένων X διάστασης D που περιέχει N δείγματα. Στόχος είναι να διασπαστεί το X σε Q κλάσεις L_q, καθεμία από τις οποίες να αντιπροσωπεύεται από μία γκαουσιανή G_q, όπου $q = 1, 2, \ldots, Q$. Στη συνέχεια περιγράφουμε δύο βασικά κριτήρια του αλγορίθμου καθώς και τα βασικά του βήματα.

8.1.1 Κριτήριο Πολυδιάστατης Κανονικότητας με βάση την Απόσταση Mahalanobis

Το κριτήριο πολυδιάστατης κανονικότητας εξετάζει αν ένα σύνολο δεδομένων X κατανέμεται σύμφωνα με την πολυδιάστατη κανονική κατανομή. Αυτό γίνεται σύμφωνα με την παρακάτω διαδικασία.

1. Εύρεση της απόστασης Mahalanobis των διανυσμάτων $\mathbf{x}_i \in X$ από το κέντρο $\overline{\mathbf{x}}$ του X
$$r_i = (\mathbf{x}_i - \overline{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}}), \quad \text{για } i = 1, \ldots, N$$
2. Ταξινόμηση των $\{r_i\}_{i=1}^N$ κατά αύξουσα σειρά και υπολογισμός της αθροιστικής συνάρτησης κατανομής της τυχαίας μεταβλητής R_i που λαμβάνει την τιμή r_i ως $\hat{F}_{R_i}(r_i) =$

i/N.

3. Αν N_{r_i} το πλήθος των διανυσμάτων x_i μέσα στην r_i-ισοπίθανη έλλειψη, τότε το N_{r_i} είναι μία διωνυμική κατανομή με παραμέτρους N και $\hat{F}_{R_i}(r_i)$. Υπολογίζονται τα $k^l_{i;\lambda}$ και $k^h_{i;\lambda}$, που είναι τα κάτω και άνω άκρα του διαστήματος εμπιστοσύνης του N_{r_i} σε 100λ% στάθμη σημαντικότητας για $i = 1, \ldots, N$ [91].

4. Η υπόθεση ότι το Q αποκλίνει από γκαουσιανή κατανομή γίνεται αποδεκτή σε επίπεδο εμπιστοσύνης 100λ% αν $\hat{F}_{R_i}(r_i) \in (\frac{k^l_{i;\lambda}}{N}, \frac{k^h_{i;\lambda}}{N})$ για τουλάχιστον $\lambda|X|$ από τις $|X|$ φορές. Το D_X είναι ένα μέτρο της μη κανονικότητας του συνόλου των δεδομένων X και μετράει πόσες φορές το $\hat{F}_{R_i}(r_i)$ βρίσκεται εκτός του διαστήματος $(\frac{k^l_{i;\lambda}}{N}, \frac{k^h_{i;\lambda}}{N})$. Αν $D_X > (1-\lambda)N$, το X δεν ακολουθεί την κανονική κατανομή με ποσοστό εμπιστοσύνης 100λ%.

8.1.2 Κριτήριο Πολυδιάστατης Κύρτωσης

Η πολυδιάστατη κύρτωση είναι μέτρο της κεντρικής τάσης ενός συνόλου διανυσμάτων και έχει βρεθεί ότι μπορεί να ανιχνεύσει αν μία ομάδα διανυσμάτων αποτελούνται από δύο ή περισσότερες γκαουσιανές με κοινά κέντρα. Η κύρτωση ορίζεται ως

$$K(Q) = \frac{1}{N}\sum_{i=1}^{N} r_i^2$$

όπου r_i είναι η απόσταση Mahalanobis.

Μεγάλη τιμή του K υποδηλώνει ότι το X προέρχεται από πλατύκυρτη κατανομή, ενώ μικρή τιμή υποδηλώνει λεπτόκυρτη κατανομή. Η λεπτόκυρτη κατανομή συμβαίνει μόνο όταν δύο ή περισσότερες γκαουσιανές μοιράζονται το ίδιο κέντρο, ενώ η πλατύκυρτη όταν η απόσταση μεταξύ των κέντρων των γκαουσιανών είναι μεγάλη.

8.1.3 Βασικά Βήματα Αλγορίθμου

1. Αρχικοποίηση: Το σύνολο δεδομένων X θεωρείται ότι αντιπροσωπεύεται από μία κλάση L_1 που ακολουθεί τη γκαουσιανή κατανομή G_1, δηλαδή $L_1 \sim G_1$, οπότε ο συνολικός αριθμός γκαουσιανών είναι ίσος με 1, δηλαδή $Q \leftarrow 1$.

   (α′) Υπολογισμός του κριτηρίου πολυδιάστατης κανονικότητας D_{L_1}

   (β′) Αν $D_{L_1} < (1-\lambda)|L_1|$, τερματισμός, αλλιώς διαίρεση της L_1 σε δύο κλάσεις, σύμφωνα με τα βήμα 4.

2. Ανάθεση κάθε διανύσματος χαρακτηριστικών $\mathbf{x}_i$, $i = 1, 2, \ldots, N$ σε μία κλάση $L_1, \ldots, L_Q$ ως εξής: Έστω w_{nq} το ποσοστό ανάθεσης του $n_{\text{οστού}}$ διανύσματος στην $q_{\text{οστή}}$ κλάση, όπως προκύπτει από τον κλασικό αλγόριθμο EM για μείγμα γκαουσιανών, και ρ_i, $i = 1, \ldots, N$ οι πραγματώσεις μίας τυχαίας μεταβλητής ομοιόμορφα κατανεμημένης στο [0,1]. Αν $\rho_i \in [\sum_{q'=1}^{q-1} w_{iq'}, \sum_{q'=1}^{q} w_{iq'}]$, τότε $\mathbf{x}_i \in L_q$.

3. Έλεγχος της υπόθεσης H_0 ότι το σύνολο των κλάσεων $L_1, \ldots, L_Q$ με γκαουσιανές κατανομές $G_1, \ldots, G_Q$ ακολουθούν το κριτήριο πολυδιάστατης κανονικότητας, $H_0 = \{L_q \sim G_q\}_{q=1}^{Q}$.

(α′) Εύρεση του $q* = arg\max_{q=1,\ldots,Q}[D_{L_q} - (1-\lambda)|L_q|]$

(β′) Αν $D_{L_{q*}} < (1-\lambda)|L_{q*}|$, τερματισμός, αλλιώς συνέχεια στο βήμα 4.

4. Διαίρεση της κλάσης L_{q*} στις κλάσεις L'_{q*} και $L_{Q'}$, όπου $Q' \leftarrow Q+1$ και υπολογισμός της πολυδιάστατης κύρτωσης $K(L_{q*})$.

   (α′) Αν $K(L_{q*})$ έχει μεγάλη τιμή, τότε το X είναι λεπτόκυρτο. Αρχικοποίηση των δύο καινούριων κέντρων στο υπάρχον κέντρο: $\mathbf{m}'_{q*} \leftarrow \mathbf{m}_{q*}$ και $\mathbf{m}'_{Q'} \leftarrow \mathbf{m}_{q*}$. Τυχαία αρχικοποίηση της διασποράς των δύο νέων κλάσεων $\mathbf{S}'_{q*}$ και $\mathbf{S}'_{Q'}$. Υποδιπλασιασμός των καινούριων ποσοστών συμμετοχής σε κάθε κλάση: $\pi'_{q*} \leftarrow \frac{|L_{q*}|}{2|X|}$ και $\pi'_{Q'} \leftarrow \frac{|L_{q*}|}{2|X|}$.

   (β′) Αν $K(L_{q*})$ έχει μικρή τιμή, τότε το X είναι πλατύκυρτο. Διαίρεση της κλάσης L_{q*} με βάση ένα διαχωριστικό υπερεπίπεδο που υπολογίζεται σε σχέση με τις οριακές κάθε άξονα στατιστικές. Το διαχωριστικό υπερεπίπεδο είναι η τιμή του διανύσματος $x_{i*} \in L_{q*}$ πάνω στον άξονα X_{d*} έτσι ώστε

$$x_{i*d*} = arg \max_{d=1,\ldots,D,\, i=1,\ldots,|L_{q*}|} F_{X_d}(x_{id}) - \hat{F}_{X_d}(x_{id})$$

   όπου $F_{X_d}(x_{id})$ είναι η θεωρητική οριακή γκαουσιανή αθροιστική συνάρτηση κατανομής με παραμέτρους την οριακή δειγματική μέση τιμή και διασπορά, ενώ $\hat{F}_{X_d}(x_{id})$ είναι η εμπειρική γκαουσιανή αθροιστική συνάρτηση κατανομής.

   Οι υπόλοιπες κλάσεις παραμένουν άθικτες.

5. Εφαρμογή κλασσικού αλγορίθμου EM για μείγμα γκαουσιανών με βάση τις κλάσεις $\{L'_q\}_{q=1}^{Q'}$ και επιστροφή στο βήμα 2.

Ο αλγόριθμος EM για μείγμα γκαουσιανών που χρησιμοποιεί την απόσταση Mahalanobis διαιρεί το σύνολο των δεδομένων σε κλάσεις L_q, $q = 1, \ldots, Q$ των οποίων το πλήθος δεν είναι απαραίτητα ίσο με το πλήθος των συναισθημάτων προς ταξινόμηση E_p, $p = 1, \ldots, P$. Για να βρούμε το ποσοστό επιτυχίας για κάθε κλάση συναισθήματος E_p, ορίζουμε M_{pq} το πλήθος των προτάσεων του συναισθήματος E_p που έχουν ταξινομηθεί στην κλάση L_q. Θεωρούμε ότι το πλήθος των προτάσεων n_q που έχει ταξινομηθεί σωστά στην κλάση q αντιστοιχεί στο συναίσθημα e_q και βρίσκονται ως

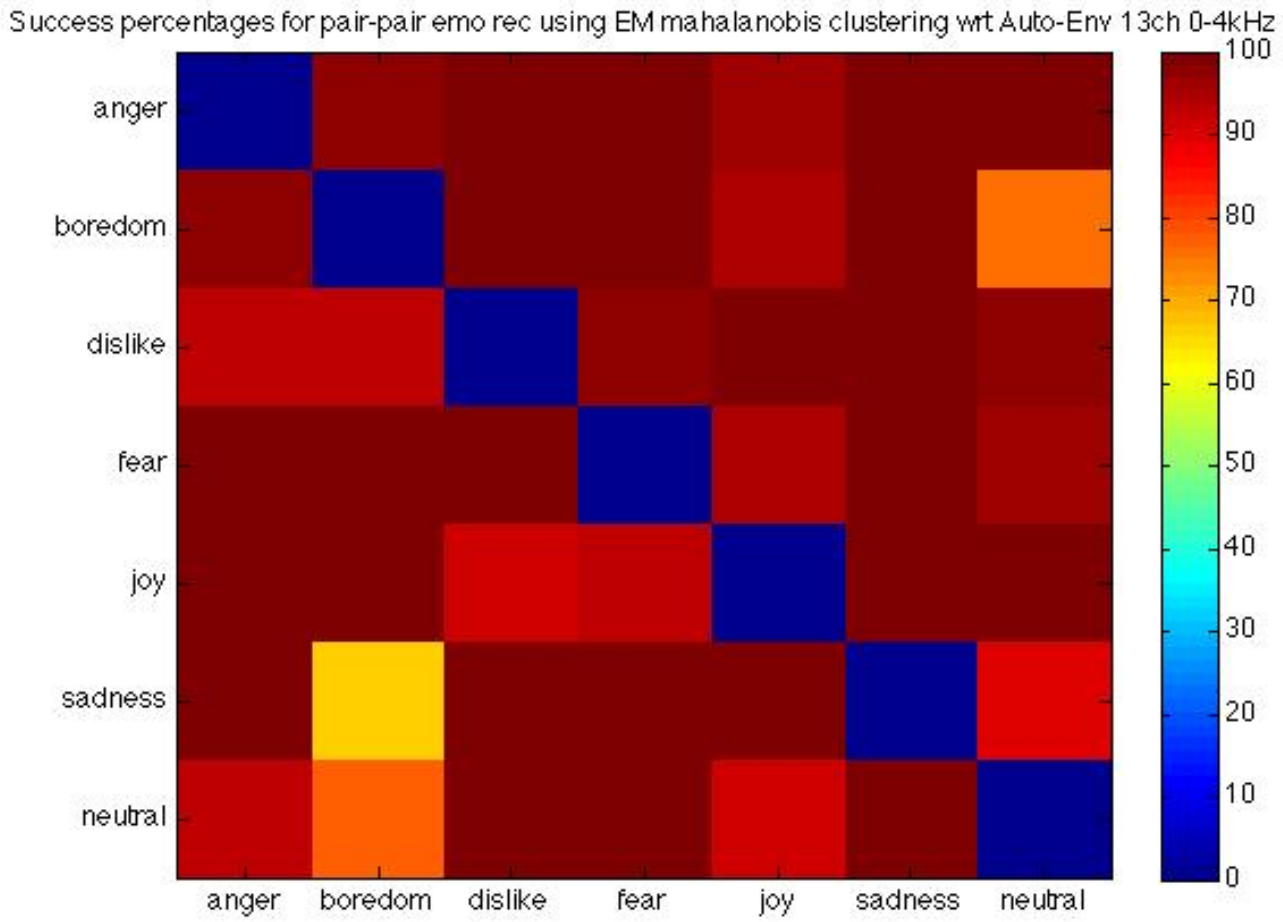
$$n_q = \max_{p=1,\ldots,P} M_{pq}, \quad e_q = arg \max_{p=1,\ldots P} M_{pq}$$

Έτσι, το συνολικό ποσοστό σωστής ταξινόμησης της κλάσης E_p των συναισθημάτων είναι

$$N_p = \frac{\sum_{e_q=p} n_q}{\sum_{q=1}^{Q} n_q}$$

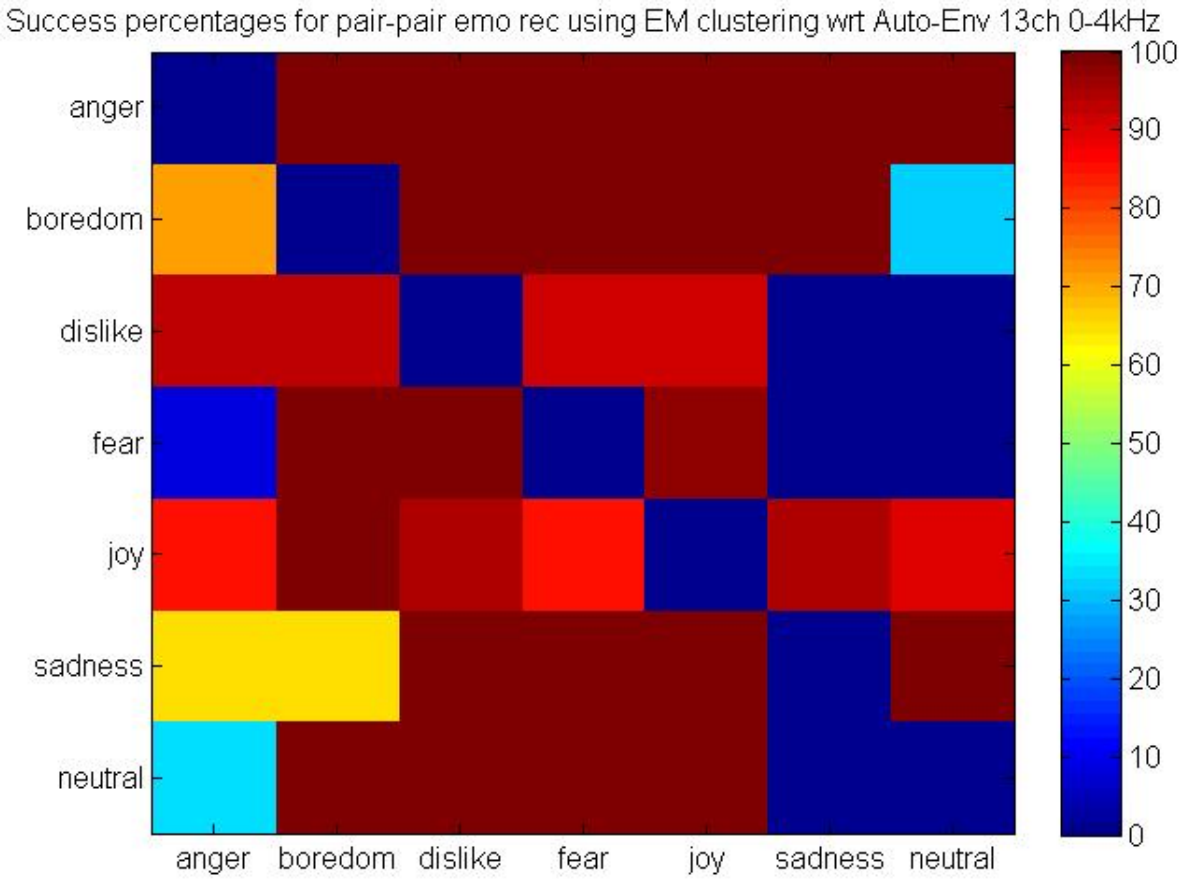
8.2 Πειραματικά Αποτελέσματα

Βλέπουμε τα ποσοστά επιτυχίας του EM σε μείγμα γκαουσιανών με τη βοήθεια της απόστασης Mahalanobis κατά την ταξινόμηση των συναισθημάτων ανά 2 με βάση το Auto-Env χαρακτηριστικό υπολογισμένο σε εύρος συχνοτήτων 0-4000Hz σε 13 κανάλια. Τα ποσοστά αυτά συγκρίνονται με τα αντίστοιχα από τον απλό EM για μείγμα γκαουσιανών και παρατηρούμε ότι είναι πολύ καλύτερα. Τα σχήματα 8.1 και 8.2 απεικονίζουν ένα τετράγωνο 7×7. Στη θέση (i, j) του τετραγώνου φαίνεται το ποσοστό επιτυχίας του συναισθήματος της κλάσης E_i, ενώ στη θέση (j, i) βρίσκεται το αντίστοιχο ποσοστό για το συναίσθημα E_j. Τα ποσοστά απεικονίζονται με βάση την κλίμακα χρωμάτων, οπότε τα τετράγωνα με κόκκινο και μπορντό χρώμα αντιστοιχούν σε μεγάλα ποσοστά αναγνώρισης, αυτά με κίτρινο και πράσινο σε μεσαία ποσοστά και αυτά με γαλάζιο και μπλε σε μικρά ποσοστά.



Σχήμα 8.1: Ποσοστά επιτυχίας αναγνώρισης συναισθημάτων ανά δύο με την EM Mahalanobis ταξινόμηση και με βάση το Auto-Env χαρακτηριστικό σε 0-4kHz και 13 κανάλια

Ο πίνακας 8.2 δείχνει το μέσο όρο των αποτελεσμάτων για διάφορους συνδυασμούς χαρακτηριστικών, που ταξινομούνται με τη μέθοδο EM για μείγμα γκαουσιανών με τη βοήθεια της απόστασης Mahalanobis. Επίσης, στο σχήμα 8.3 βλέπουμε το ποσοστό επιτυχίας για 6

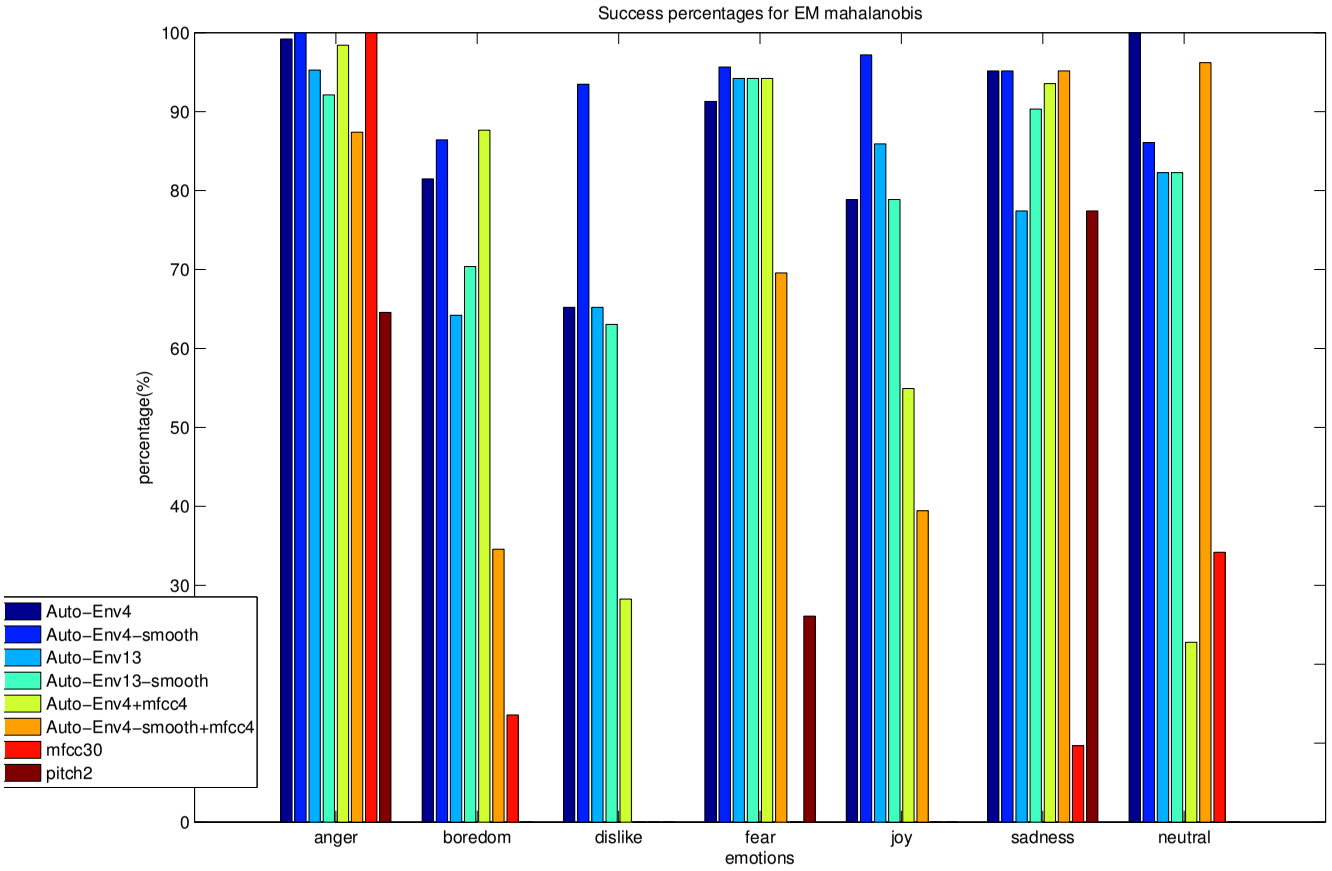


Σχήμα 8.2: Ποσοστά επιτυχίας αναγνώρισης συναισθημάτων ανά δύο με την απλή EM ταξινόμηση και με βάση το Auto-Env χαρακτηριστικό σε 0-4kHz και 13 κανάλια

συναισθήματα και το ουδέτερο. Το Auto-Env επιτυγχάνει καλύτερα αποτελέσματα από όλα τα υπόλοιπα χαρακτηριστικά, ενώ η μέθοδος EM για μείγμα γκαουσιανών και με τη βοήθεια της απόστασης Mahalanobis είναι πιο αποδοτική για χαρακτηριστικά μικρής διάστασης. Ειδικότερα, το Auto-Env που έχει υπολογιστεί σε 4 κανάλια είναι πολύ πιο αποτελεσματικό από το ίδιο χαρακτηριστικό σε 13 κανάλια, με ποσοστά επιτυχίας 87.32% και 81.60% αντίστοιχα. Επίσης, η ομαλοποιημένη εκδοχή του Auto-Env βελτιώνει την αναγνώριση, δίνοντας ποσοστό 93.42% για 4 κανάλια, που επιφέρει 6.1% απόλυτη αύξηση.

EM για μείγμα γκαουσιανών με τη βοήθεια της απόστασης Mahalanobis.		
	Ποσοστά επιτυχίας (%)	
Περιγραφή χαρακτηριστικών	Μέσος όρος	Απόκλιση
Auto-Env (4ch,0-4kHz)	87.32	12.74
Auto-Env smoothed(4ch,0-4kHz)	93.42	5.30
Auto-Env(13ch,0-4kHz)	80.64	12.56
Auto-Env smoothed(13ch,0-4kHz)	81.60	11.72
Auto-Env smoothed(4ch,0-4kHz)+pitch	82.34	17.04
Auto-Env+mfcc(4ch,0-4kHz)	68.54	32.77
Ampl-area smoothed(4ch,0-4kHz)	44.33	27.15
F(4ch,0-4kHz)	41.56	27.02
Ampl-kurt smoothed(4ch,0-4kHz)	32.69	25.96
Ampl-kurt smoothed(4ch,0-4kHz)	31.41	39.58
mfcc (30ch,0-4kHz)	22.49	36.30

Πίνακας 8.1: Ποσοστά επιτυχίας συνδυασμών ESA, pitch και MFCC χαρακτηριστικών που έχουν υποστεί LDA με τη μέθοδο EM mahalanobis για μείγμα γκαουσιανών



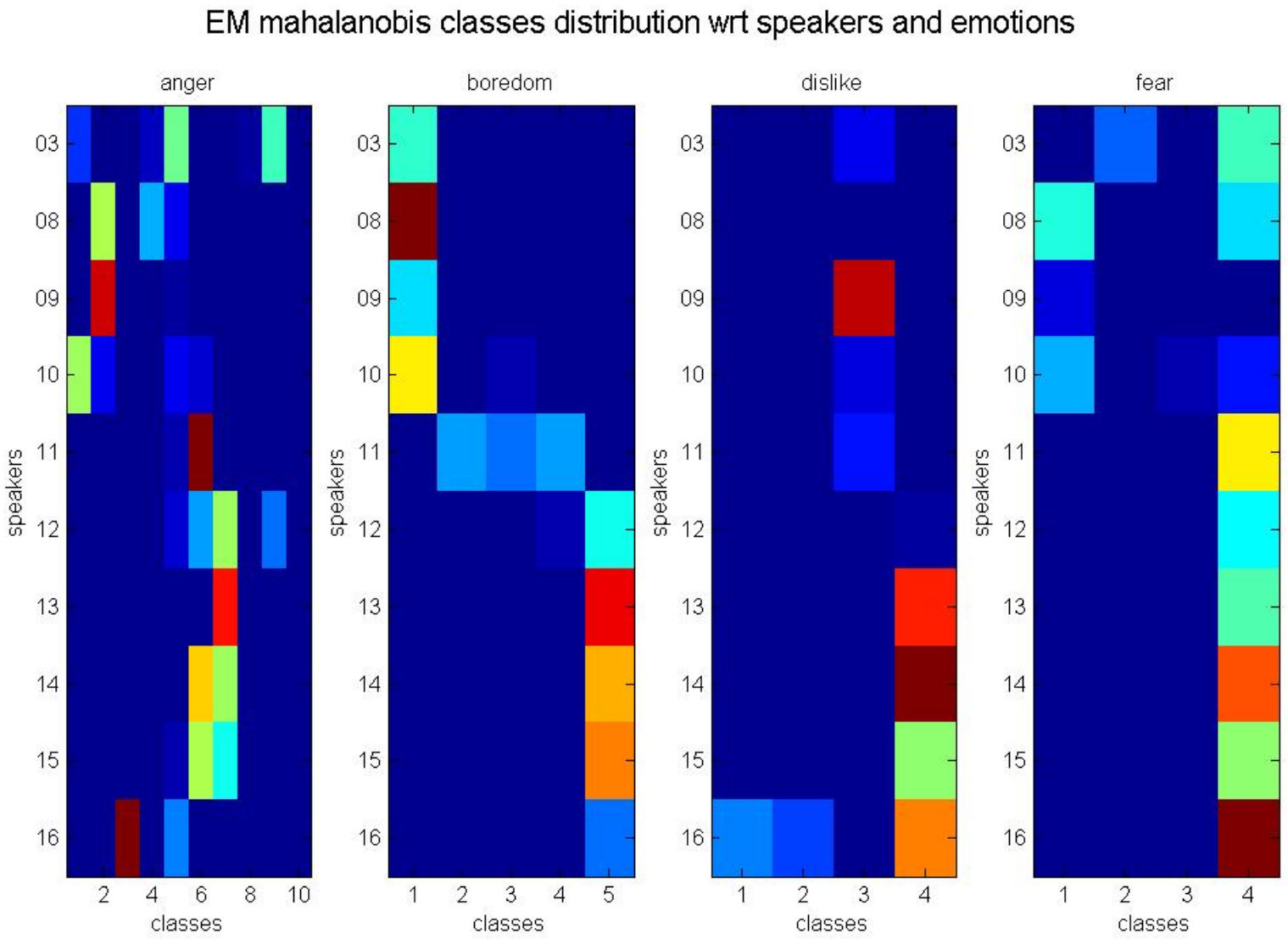
Σχήμα 8.3: Ποσοστά επιτυχίας αναγνώρισης 6 συναισθημάτων και του ουδέτερου με τη μέθοδο EM για μείγμα γκαουσιανών και με τη βοήθεια της απόστασης Mahalanobis

8.3 Εξάρτηση από τον Ομιλητή και την Πρόταση

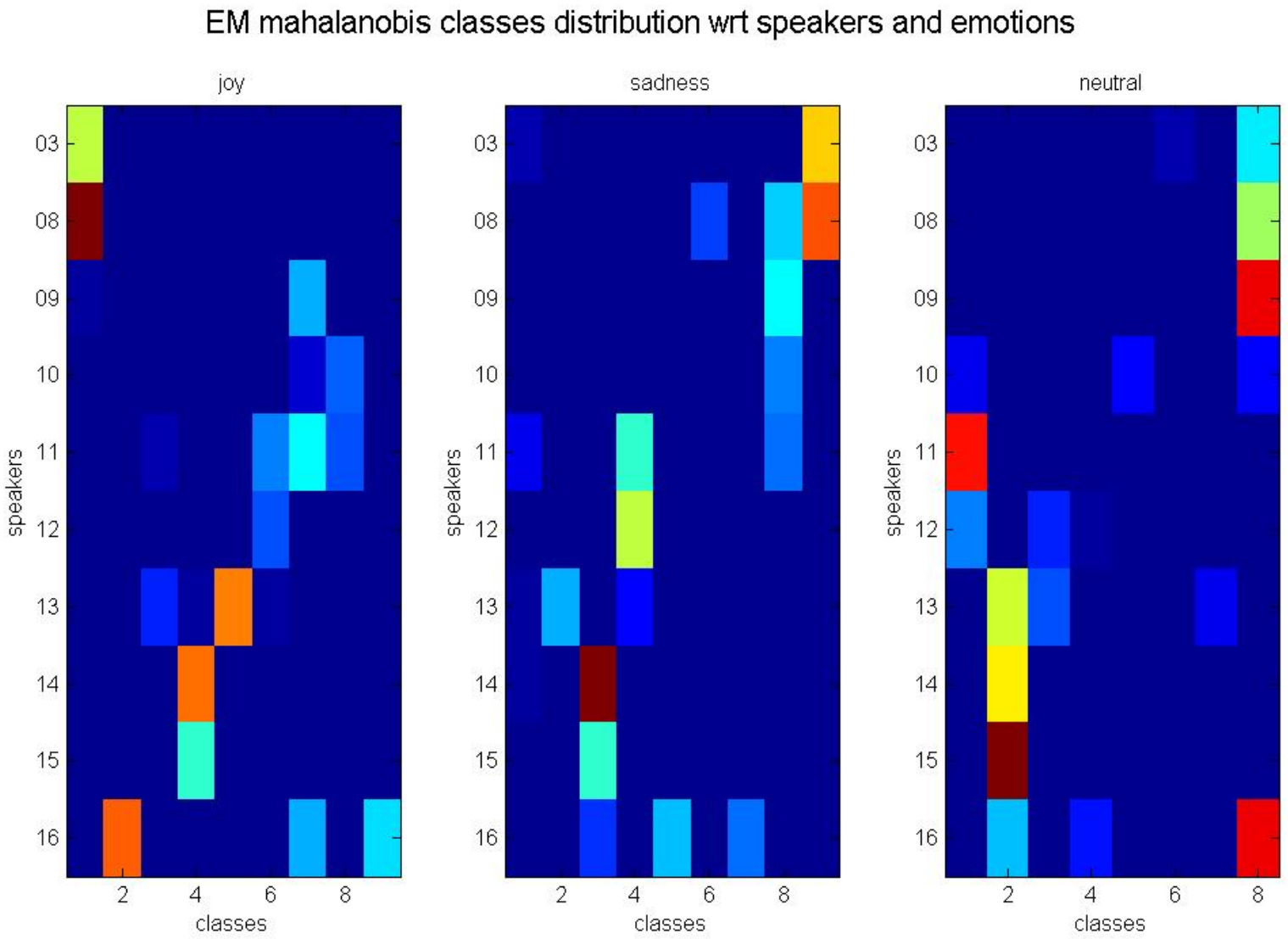
Το μείγμα γκαουσιανών που βρίσκεται με τη βοήθεια της απόστασης Mahalanobis αποτελείται συνήθως από πλήθος γκαουσιανών μεγαλύτερο του αριθμού των συναισθημάτων υπό ταξινόμηση. Για το λόγο αυτό, εξετάζουμε τις προτάσεις που κατηγοριοποιούνται στην ίδια κλάση με βάση τον αλγόριθμο αυτό. Στη βάση δεδομένων Berlin database of emotional speech 10 ομιλητές έχουν εκφωνήσει 10 διαφορετικές προτάσεις με 6 συναισθήματα και το ουδέτερο. Επιχειρούμε να δούμε κατά πόσο το καλύτερο μέχρι τώρα χαρακτηριστικό, που είναι το Auto-Env υπολογισμένο σε 13 κανάλια, κατηγοριοποιεί τα συναισθήματα ανεξαρτήτως πρότασης και ομιλητή.

Στα σχήματα 8.4 και 8.5 παρατηρούμε την κατανομή των ομιλητών σε συνάρτηση με την κλάση ταξινόμησης του EM για μείγμα γκαουσιανών mahalanobis αλγορίθμου για κάθε συναίσθημα. Στον κατακόρυφο άξονα παρουσιάζεται ο κωδικός αριθμός κάθε ομιλητή, ενώ στον οριζόντιο βρίσκεται η κλάση που προκύπτει από τον EM. Η πλειονότητα των ομιλητών ταξινομούνται στην ίδια κλάση στο συναίσθημα του φόβου, καθώς στο αντίστοιχο σχήμα παρατηρούμε μία κατακόρυφη γραμμή με πράσινο, κίτρινο, κόκκινο και μπορντό χρώμα στην 4η κλάση. Αυτό σημαίνει ότι οι περισσότεροι ομιλητές εκφώνησαν τις προτάσεις φόβου με τον ίδιο τρόπο, που μπορεί να διακριθεί αποτελεσματικά από το Auto-Env για 4 κανάλια. Αντίθετα, στο θυμό και στη χαρά οι ομιλητές έχουν κατηγοριοποιηθεί σε αρκετές κλάσεις. Αυτό δείχνει ότι οι ομιλητές δεν εκφράζουν με ομοιόμορφο τρόπο τα συναισθήματα αυτά ή ότι το Auto-Env δεν είναι απόλυτα διαχωρίσιμο χαρακτηριστικό για αυτά. Είναι επίσης αξιοσημείωτο ότι για πολλά συναισθήματα υπάρχουν ομιλητές που κατηγοριοποιούνται σε ίδιες ομάδες μεταξύ τους. Για παράδειγμα, οι ομιλητές με κωδικό 13, 14, 15 και 16 κατηγοριοποιούνται στην ίδια κλάση για την πλήξη, την απαρέσκεια και το φόβο. Παρόμοια παρατήρηση μπορεί να γίνει για τους ομιλητές 3 και 8 για την πλήξη, το φόβο, τη χαρά και το ουδέτερο. Από τις πληροφορίες που δίνονται στη βάση δεδομένων, γνωρίζουμε ότι οι ομιλητές 13, 14 και 16 είναι γυναίκες παρόμοιας ηλικίας (32, 35 και 31 χρονών αντίστοιχα), οπότε αυτή η παρατήρηση είναι ενδεχομένως μία ένδειξη για την εξάρτηση της έκφρασης του συναισθήματος από το φύλο.

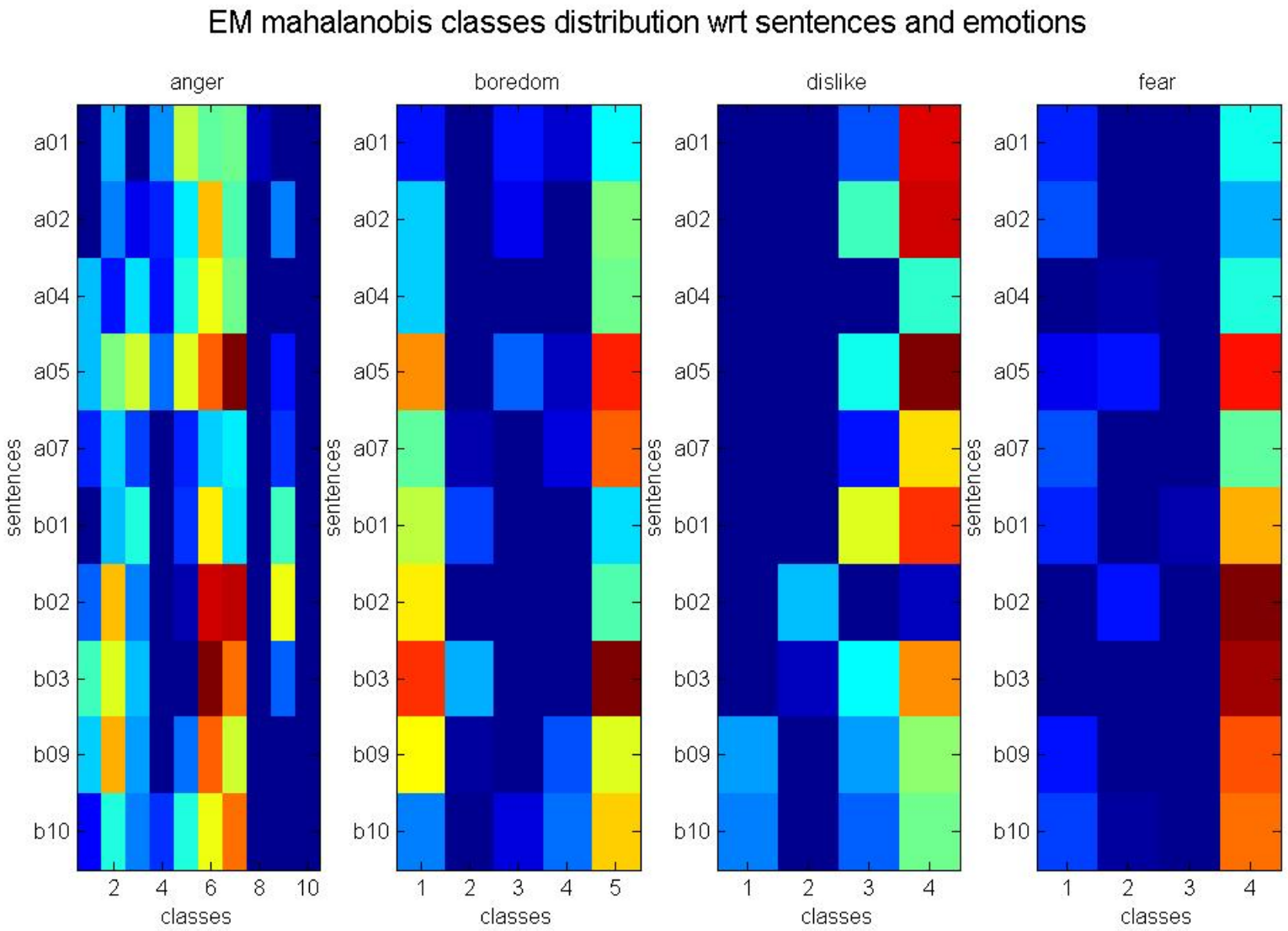
Τέλος, επιχειρούμε να διαπιστώσουμε κατά πόσο το περιεχόμενο των προτάσεων επηρεάζει την ταξινόμηση των συναισθημάτων με βάση τον EM mahalanobis αλγόριθμο και το Auto-Env χαρακτηριστικό. Στα σχήματα 8.6 και 8.7 παρατηρούμε την κατανομή των ομιλητών σε συνάρτηση με την κλάση ταξινόμησης. Καθεμία πρόταση μπορεί να κατηγοριοποιηθεί σε περισσότερες από μία κλάσεις για κάθε συναίσθημα, πράγμα που σημαίνει ότι το σύστημα ταξινόμησης είναι ανεξάρτητο του περιεχομενου των προτάσεων.



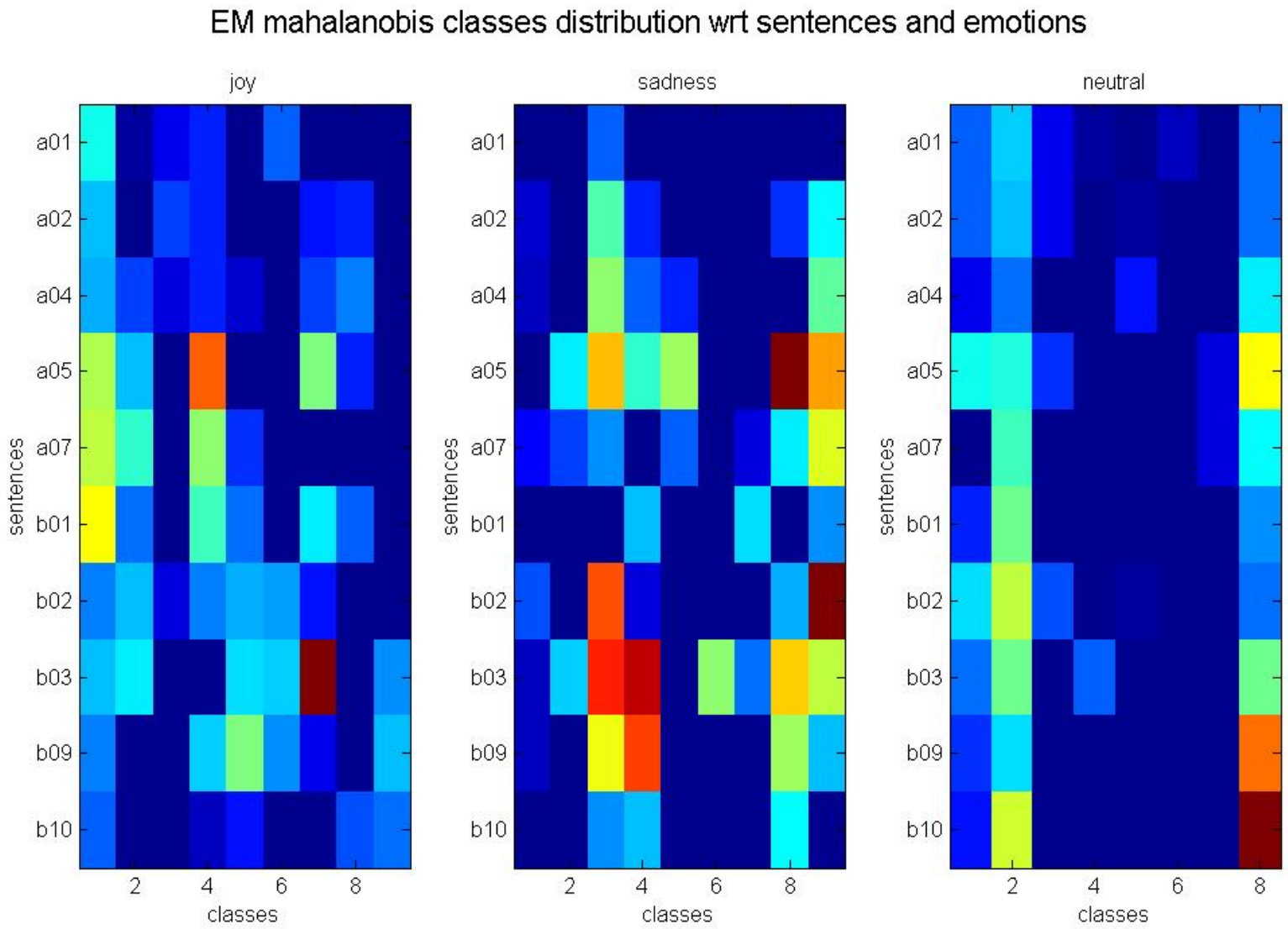
Σχήμα 8.4: Κατανομή ομιλητών σε συνάρτηση με την κλάση ταξινόμησης του EM για μείγμα γκαουσιανών mahalanobis αλγορίθμου με βάση το Auto-Env χαρακτηριστικό για κάθε συναίσθημα.



Σχήμα 8.5: Κατανομή ομιλητών σε συνάρτηση με την κλάση ταξινόμησης του EM για μείγμα γκαουσιανών mahalanobis αλγορίθμου με βάση το Auto-Env χαρακτηριστικό για κάθε συναίσθημα.



Σχήμα 8.6: Κατανομή προτάσεων σε συνάρτηση με την κλάση ταξινόμησης του EM για μείγμα γκαουσιανών mahalanobis αλγορίθμου με βάση το Auto-Env χαρακτηριστικό για κάθε συναίσθημα.



Σχήμα 8.7: Κατανομή προτάσεων σε συνάρτηση με την κλάση ταξινόμησης του EM για μείγμα γκαουσιανών mahalanobis αλγορίθμου με βάση το Auto-Env χαρακτηριστικό για κάθε συναίσθημα.

Κεφάλαιο 9

Δημιουργία και Έλεγχος Βάσης Δεδομένων Αιγινήτειου Νοσοκομείου

Στόχος είναι η κατασκευή βάσης προφορικού λόγου συναισθηματικής έκφρασης. Επιλέξαμε 5 βασικά συναισθήματα: θυμός, φόβος, χαρά, λύπη και ουδέτερο. Για λόγους απλότητας και καλύτερου ελέγχου του πειράματος, θεωρήσαμε πιο αποδοτικό να καλέσουμε άτομα παρόμοιας ηλικίας να εκφωνήσουν προκαθορισμένες προτάσεις με τα 5 συναισθήματα. Για κάθε συναίσθημα κατασκευάσαμε 8 προτάσεις με περιεχόμενο αντιπροσωπευτικό για το συναίσθημα αυτό, έτσι ώστε να μπορούν τα άτομα να εκφραστούν με όσο το δυνατό μεγαλύτερη φυσικότητα.

9.1 Οργάνωση Εργαστηριακού Εξοπλισμού

Οι καταγραφές έγιναν σε έναν ειδικά μονωμένο χώρο εργαστηρίου, που φαίνεται στην εικόνα 9.1, ώστε να έχει όσο το δυνατό λιγότερο θόρυβο και λιγότερες ανακλάσεις. Ο εξοπλισμός που χρησιμοποιήθηκε είναι:

1. κάψα μικροφώνου Sennheiser ME67
2. τροφοδοτικό μικροφώνου Senheiser K6
3. προενισχυτής Digimax FS
4. κάρτα ήχου Digital Konnekt x32
5. λογισμικό καταγραφής Cubase Essential 5
6. μηχάνημα καταγραφής εικόνας Digital Video Recorder 4008

Δύο εικόνες του εξοπλισμού φαίνονται στο σχήμα 9.2.

9.2 Δημιουγία Προτάσεων προς Εκφώνηση για την Έκφραση Συναισθημάτων

Για να ελέγξουμε την εγκυρότητα κάθε πρότασης, ζητήσαμε από 40 φοιτητές (24 άντρες και 16 γυναίκες) να αξιολογήσουν τις 40 προτάσεις που δημιουργήσαμε, δηλαδή να σημειώσουν ποια συναισθηματική κατάσταση θεωρούν ότι εκφράζουν. Το 65% των φοιτητών είναι προπτυχιακοί, το 20% είναι μεταπτυχιακοί, ενώ το 15% έχουν τελειώσει και τις μεταπτυχακές σπουδές. Στο σχήμα 9.3 φαίνεται ένα ιστόγραμμα με τις ηλικίες αυτών που συμμετέχουν στην έρευνα.

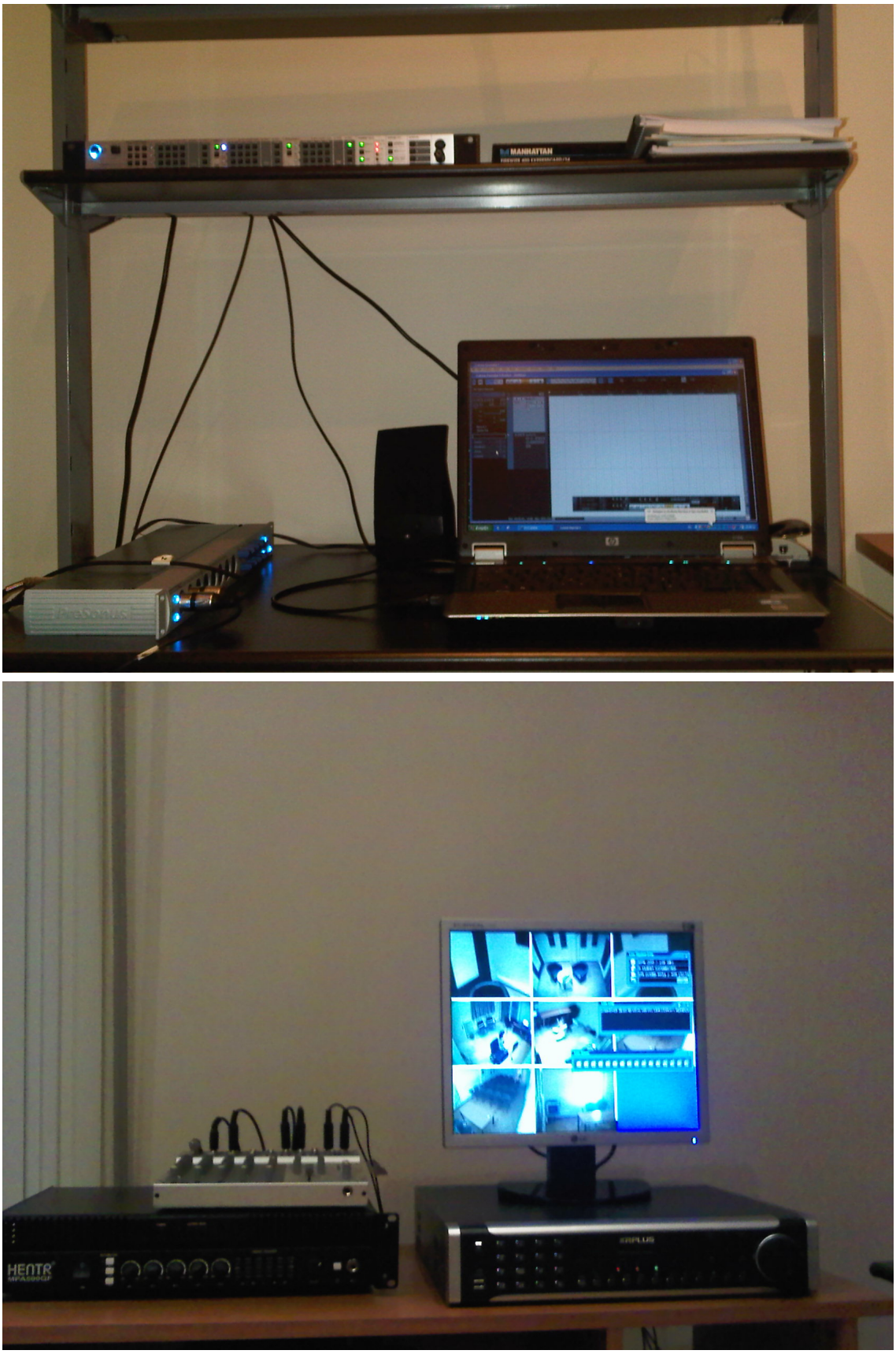
Τα αποτελέσματα για καθεμία από τις προτάσεις αυτές φαίνονται στους παρακάτω πίνακες. Για την εγκυρότητα της έρευνας θα κρατήσουμε τις προτάσεις στις οποίες πάνω από το 85% των ερωτηθέντων απάντησαν σωστά. Έτσι, θεωρούμε αποτελεσματικές τις 7 από τις 8 προτάσεις. Στο σχήμα 9.5 παρουσιάζεται ο μέσος όρος αναγνώρισης κάθε συναισθήματος στην έρευνα που έγινε, όπως προκύπτει από τις 7 προτάσεις που τελικά θα χρησιμοποιήσουμε.

Οι 7 από τις 8 προτάσεις που αναγνωρίστηκαν με επιτυχία 85% και άνω, καταγράφηκαν στο εργαστήριο με τον εξοπλισμό που περιγράψαμε. Στις καταγραφές συμμετείχαν 16 άτομα (φοιτητές ηλικίας 20-30 χρονών), 11 άντρες και 5 γυναίκες. Συλλέχθηκαν συνολικά 556 εκφωνήσεις: 111 θυμού, 111 φόβου, 111 χαράς, 111 ουδέτερου και 112 λύπης. Στον πίνακα 9.1 φαίνονται συγκεντρωτικά οι προδιαγραφές της βάσης δεδομένων που κατασκευάσαμε και στις εικόνες 9.4 φαίνονται κάποια στιγμιότυπα από τις καταγραφές.

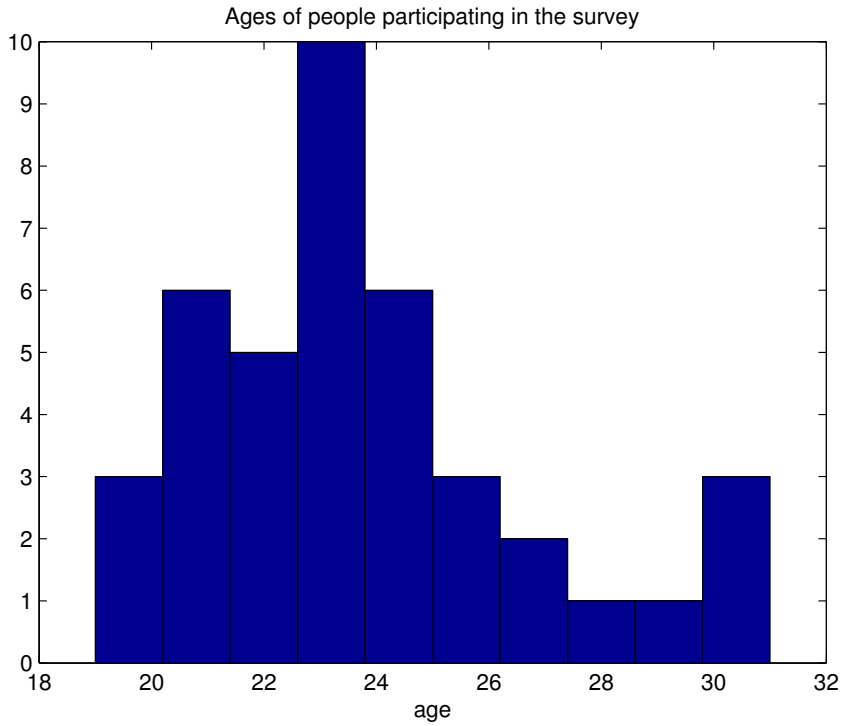
Αριθμός ομιλητών						16			
Προτασεις ανά ομιλητή						7			
Επαναλήψεις ανά πρόταση						1			
Συνολικός αριθμός εκφωνήσεων						556			
Αριθμός εκφωνήσεων ανά συναίσθημα									
θυμός	111	**φόβος**	111	**χαρά**	111	**λύπη**	112	**ουδ**	111
Συνολική χρονική διάρκεια βάσης						22min 30sec			
Μέση χρονική διάρκεια ανά πρόταση (sec)									
A01	3.25	**F01**	1.76	**J01**	1.30	**S01**	1.89	**N01**	3.47
A02	1.88	**F02**	1.59	**J02**	2.65	**S02**	2.79	**N02**	2.75
A03	2.12	**F03**	3.45	**J03**	1.81	**S03**	1.86	**N03**	2.27
A04	2.00	**F04**	3.07	**J04**	2.09	**S04**	2.01	**N04**	1.74
A05	1.81	**F05**	3.34	**J05**	1.20	**S05**	2.21	**N05**	4.03
A06	1.91	**F06**	3.77	**J06**	1.68	**S06**	2.74	**N06**	2.44
A07	1.81	**F07**	3.08	**J07**	3.28	**S07**	2.79	**N07**	3.00
Συχνότητα δειγματοληψίας ήχου						44.1kHz			
Κατεύθυνση καταγραφής μικροφώνου						υπερκαρδιωειδής λοβός			
Ανάλυση εικόνας						720×480pixels			
Συχνότητα πλαισίου εικόνας						25frames/sec			
Μέγεθος φακού κάμερας						6-15mm			

Πίνακας 9.1: Συγκεντρωτικός πίνακας προδιαγραφών βασης δεδομένων Αιγινήτειου Νοσοκομείου.

Σχήμα 9.1: Δωμάτιο καταγραφής Αιγινήτειου Νοσοκομείου.



Σχήμα 9.2: Εξοπλισμός καταγραφής Αιγινήτειου Νοσοκομείου.

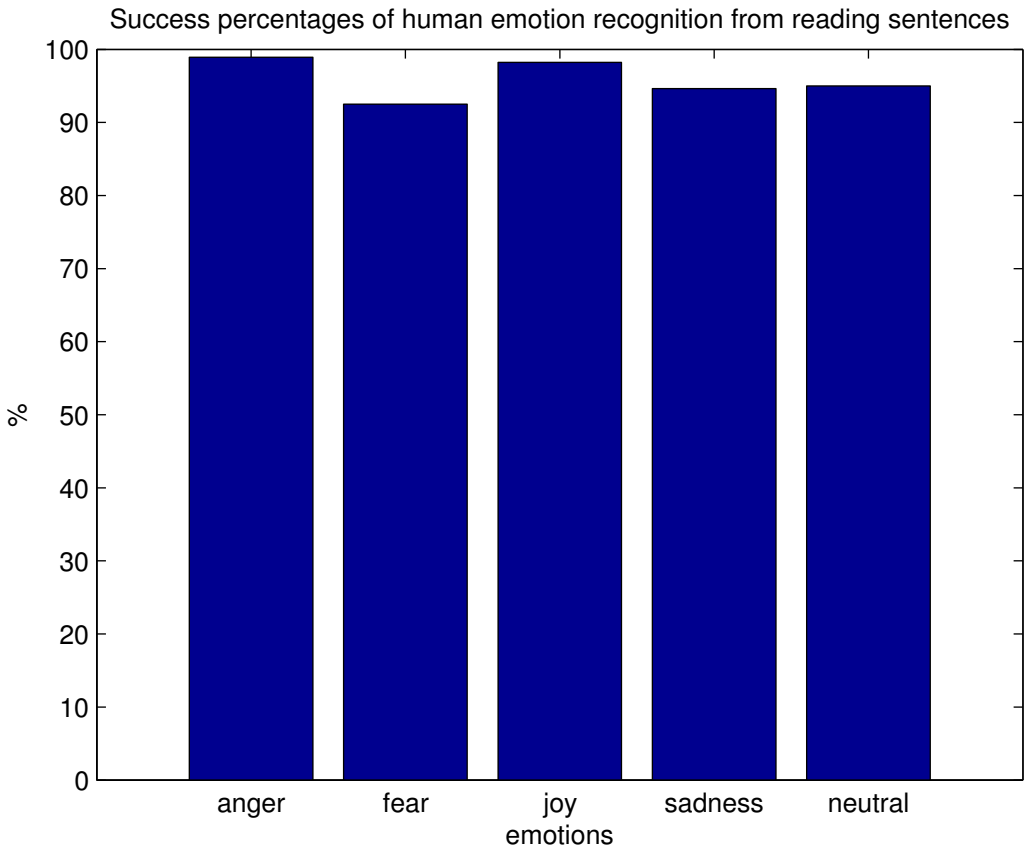


Σχήμα 9.3: Ιστόγραμμα ηλικιών ανθρώπων που συμμετέχουν στην έρευνα

Σχήμα 9.4: Στιγμιότυπα από τις καταγραφές.

	Φράση	Θυμός	Φόβος	Χαρά	Λύπη	Ουδέτερο
Θυμος	A01)Αν δε σταματήσει αμέσως αυτή η μουσική, θα φέρω την αστυνομία!	**40**	0	0	0	0
	A02)Πώς τολμάς να πειράζεις τα πράγματά μου!	**40**	0	0	0	0
	A03)Πολύ με εκνευρίζεις με την ασυνέπειά σου!	**40**	0	0	0	0
	A04)Φύγε τώρα αμέσως από το δωμάτιό μου!	**40**	0	0	0	0
	A05)Μη μου υψώνεις τον τόνο της φωνής σου!	**40**	0	0	0	0
	A06)Η συμπεριφορά σου είναι απαράδεκτη!	**39**	0	0	1	0
	A07)Πάλι δεν έβαλες τα ρούχα στη θέση τους!	**38**	0	0	0	2
	Αμάν πια! Έρχεσαι πάντα αργοπορημένος!	**34**	0	5	0	1
Φόβος	F01)Ακούω ύποπτο θόρυβο στην πόρτα.	0	**40**	0	0	0
	F02)Μη με χτυπήσεις, σε παρακαλώ.	0	**40**	0	0	0
	F03)Πω πω, δεν μπορώ να σταματήσω το αμάξι στην ανεξέλεγκτη πορεία του.	0	**39**	0	0	1
	F04)Λες να με σταματήσει η αστυνομία που υπερβαίνω το όριο ταχύτητας;	0	**36**	0	0	4
	F05)Δεν τολμάω να μπω στη θάλασσα, γιατί είναι πολύ βαθιά τα νερά για μένα.	1	**35**	0	1	3
	F06)Ωχ, πολύ άγριος μου φαίνεται αυτός ο σκύλος, ασφαλώς δαγκώνει.	1	**35**	0	0	4
	F07)Δεν ξέρω αν θα μπορέσω να βρω καινούρια δουλειά μ΄ αυτή την ανεργία.	0	**34**	0	4	2
	Δε θα ήθελα να κάνω το εμβόλιο, αφού δεν ξέρω τι παρενέργειες θα έχει.	1	**28**	0	0	11
Χαρά	J01)Αυτά είναι τέλεια νέα!	0	0	**40**	0	0
	J02)Το δώρο που μου χάρισες είναι ακριβώς αυτό που ήθελα!	0	0	**40**	0	0
	J03)Βγήκαν οι βαθμοί και πήρα άριστα!	0	0	**40**	0	0
	J04)Επιτέλους φεύγω αύριο για διακοπές!	0	0	**40**	0	0
	J05)Κέρδισα το στοίχημα!	0	0	**39**	0	1

	Φράση	Θυμός	Φόβος	Χαρά	Λύπη	Ουδέτερο
Χαρά	J06)Ηταν πολύ μεγάλη αυτή η νίκη!	0	0	**38**	0	2
	J07)Ευτυχώς το ερχόμενο σαββατοκύριακο θα μπορέσω να δω τους δικούς μου!	0	0	**38**	0	2
	Συγχαρητήρια για την επιτυχία σου!	5	0	**34**	0	1
Λύπη	S01)Πέθανε ο σκύλος που αγαπούσα πολύ.	0	0	0	**40**	0
	S02)Θα μου λέιψει η γιαγιά μου κι ας ήτανε πολύ ηλικιωμένη.	0	0	0	**40**	0
	S03)Τι κρίμα που έγινε αυτή η καταστροφή.	0	0	0	**39**	1
	S04)Δυστυχώς η καρδιά του τον πρόδωσε.	1	0	0	**38**	1
	S05)Τι κρίμα που δεν μπορώ να αποφύγω το χωρισμό.	0	0	1	**38**	1
	S06)Το πείραμα απέτυχε, πάει χαμένος ο κόπος μας.	1	1	0	**36**	2
	S07)Κόπηκα στις εξετάσεις, ενώ ήταν βέβαιο ότι θα περάσω.	6	0	0	**34**	0
	Λυπάμαι που δεν μπορούμε να τον μεταπείσουμε.	0	0	0	**33**	7
Ουδέτερο	N01)Τα κουφώματα από αλουμίνιο είναι πιο ανθεκτικά από εκείνα με ξύλο.	0	0	0	0	**40**
	N02)Σήμερα θα πάω στο Σούπερ Μάρκετ, γιατί αύριο είναι αργία.	0	0	1	0	**39**
	N03)Η σκουριά καταστρέφει σταδιακά τα σίδερα.	0	0	0	2	**38**
	N04)Τα βιβλία χρειάζονται ξεσκόνισμα.	1	0	0	1	**38**
	N05)Το επάγγελμα του ηλεκτρολόγου είναι πιο προσοδοφόρο από το επάγγελμα του υδραυλικού.	0	0	3	0	**37**
	N06)Το μαύρο ψωμί είναι πιο υγιεινό από το άσπρο.	0	1	0	2	**37**
	N07)Το σπίτι ήταν πολλές μέρες κλειστό και χρειάζεται καθάρισμα.	1	0	0	2	**37**
	Χρησιμοποιώντας λαμπτήρες νέας γενιάς, εξοικονομούμε ενέργεια.	0	0	10	0	**30**



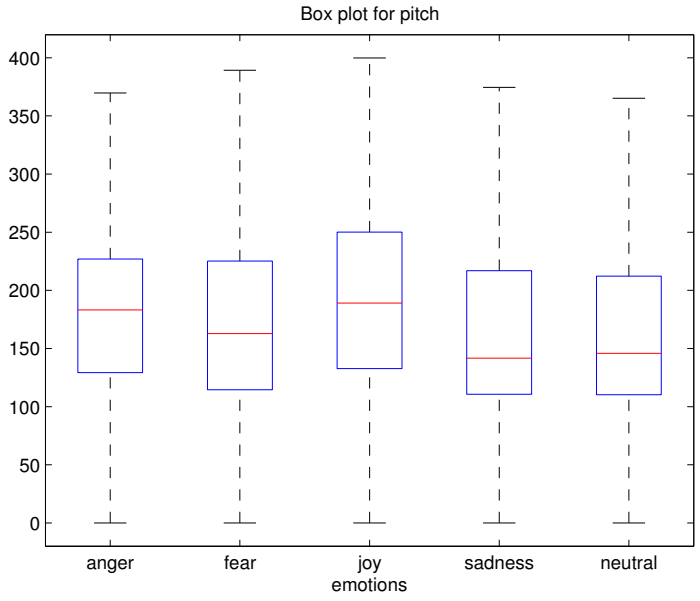
Σχήμα 9.5: Ποσοστά επιτυχίας ανθρώπινης αναγνώρισης για κάθε συναίσθημα.

9.3 Μελέτη Χαρακτηριστικών στη Βάση Δεδομένων Αιγινήτειου Νοσοκομείου

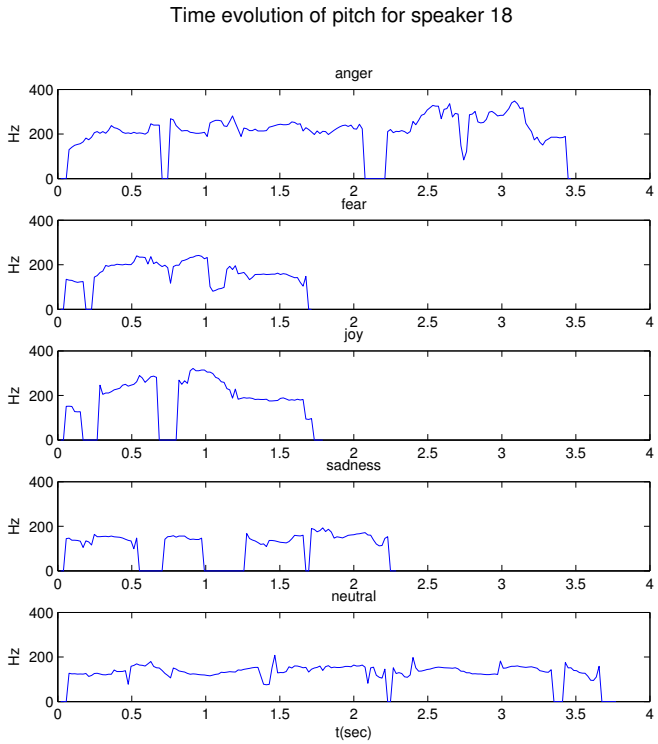
Μελετάμε τα χαρακτηριστικά που προκύπτουν από τις προτάσεις της καινούριας βάσης δεδομένων.

Στο σχήμα 9.6 φαίνεται το box plot του pitch για όλες τις προτάσεις της βάσης δεδομένων. Η μορφή του σχήματος συμφωνεί με το αντίστοιχο σχήμα για τη βάση δεδομένων Berlin DB. Όπως έχουμε προαναφέρει, ο θυμός και η χαρά έχουν μεγαλύτερες τιμές pitch απ'οτι ο φόβος και η λύπη, πράγμα που φαίνεται από την κόκκινη γραμμή της διαμέσου. Αυτό επιβεβαιώνεται και στο σχήμα 9.7, όπου παρουσιάζεται η χρονική εξέλιξη του μεγέθους αυτού για ορισμένες προτάσεις εκφρασμένες με διαφορετικά συναισθήματα από τον ίδιο ομιλητή. Βλέπουμε ότι ο θυμός και η χαρά εμφανίζουν και πιο απότομες μεταβολές κατά τη διάρκεια εκφοράς μίας πρότασης, σε αντίθεση με το ουδέτερο συναίσθημα, που έχει πιο σταθερό pitch.

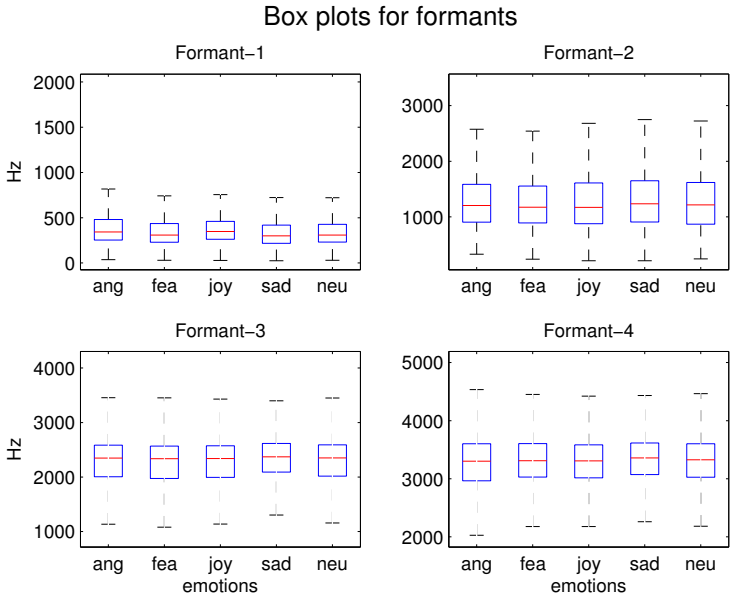
Στο σχήμα 9.8 φαίνονται τα box plots των 4 πρώτων formants για όλες τις προτάσεις. Ενώ στη βάση δεδομένων Berlin DB υπήρχαν πιο πολλλές διαφορές μεταξύ των συναισθημάτων, στη βάση του Αιγινήτειου οι πιο πολλές διαοφορές υπάρχουν μόνο στο 1ο formant. Βλέπουμε ότι οι τιμές διαμέσου του θυμού και της χαράς στο 1ο formant είναι ανεπαίσθητα πιο υψηλές από τις αντίστοιχες για το φόβο και τη λύπη, πράγμα που συμφωνεί με τις παρατηρήσεις στη βάση Berlin DB. Παρόλα αυτά η λύπη δε φαίνεται να έχει το μεγάλο εύρος τιμών 1ου formant που παρατηρήσαμε στη βάση Berlin DB.



Σχήμα 9.6: Box plots του pitch για κάθε συναίσθημα σε όλες τις προτάσεις.



Σχήμα 9.7: Χρονική εξέλιξη του pitch για μία πρόταση κάθε συναισθήματος στον ομιλητή 18.

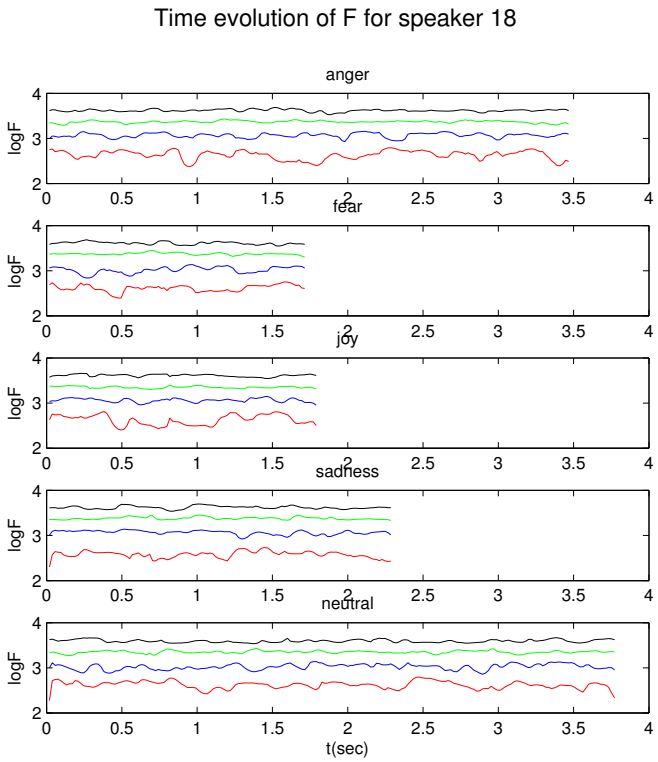


Σχήμα 9.8: Box plots των 4 πρώτων formants για κάθε συναίσθημα σε όλες τις προτάσεις.

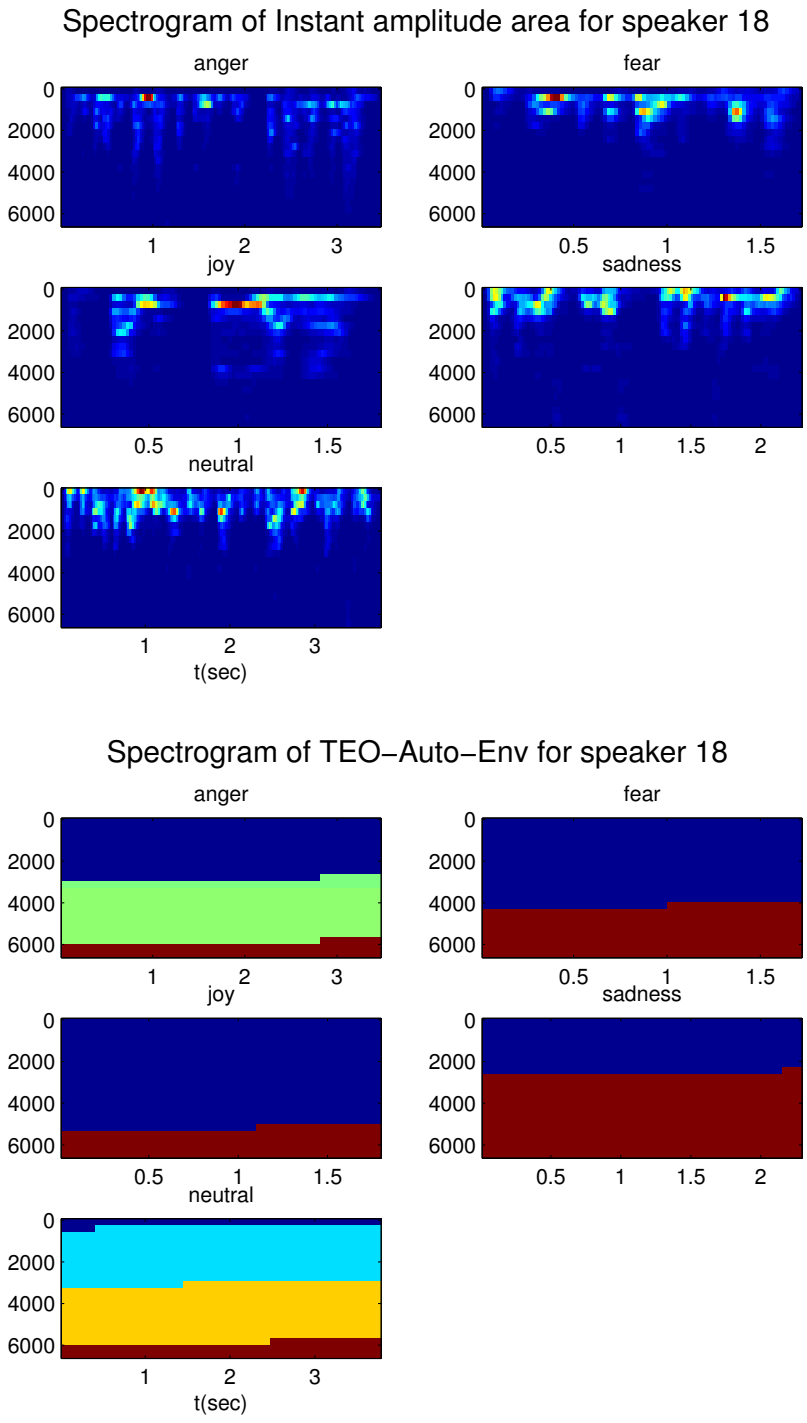
Στη συνέχεια, μελετάμε τα AM-FM χαρακτηριστικά διαμόρφωσης. Όπως είδαμε στα προηγούμενα κεφάλαια, δύο από τα πιο ισχυρά χαρακτηριστικά είναι το εμβαδόν του στιγμιαίου πλάτους (Ampl-area) και του εμβαδόν της αυτοσυσχέτισης αυτού (Auto-Env). Στο σχήμα 9.10 βλέπουμε τα φασματογράμματα των μεγεθών αυτών για τυχαίες προτάσεις σε όλα τα συναισθήματα εκφρασμένες από τον ομιλητή 18. Όπως και στη βάση δεδομένων Berlin DB, ο θυμός έχει διάσπαρτες υψηλές τιμές του Ampl-mean σε όλο το εύρος συχνοτήτων. Η χαρά εμφανίζει πιο υψηλές τιμές από το θυμό, ενώ η λύπη έχει περιορισμένες τις υψηλές τιμές Ampl-mean στις χαμηλές συχνότητες. Όσον αφορά στο TEO-Auto-Env παρατηρούμε ότι τα συναισθήματα εμφανίζουν μεγάλες διαφορές, πράγμα που θα φανεί και στη συνέχεια με τα μεγάλα ποσοστά ταξινόμησης για το χαρακτηριστικό αυτό.

Μεγάλης σημασίας είναι επίσης και τα χαρακτηριστικα στιγμιαίας συχνότητας από τα AM-FM χαρακτηριστικά διαμόρφωσης. Στο σχήμα 9.11 φαίνονται τα box plots των F και B, όπου δε διακρίνουμε πολλές διαφορές. Όπως ίσχυε και στη βάση δεδομένων Berlin DB, οι πιο μεγάλες διαφορές βρίσκονται στο 1ο κανάλι για το σταθμισμένο μέσο όρο στιγμιαίας συχνότητας. Στο σχήμα 9.9 απεικονίζεται η χρονική εξέλιξη μεμονωμένων προτάσεων του F. για τον ομιλητή 18. Πιο μεγάλες μεταβολές του μεγέθους αυτού έχουμε στο θυμό και στη χαρά, ενώ μικρότερες μεταβολές υπάρχουν στη λύπη και στο ουδέτερο συναίσθημα.

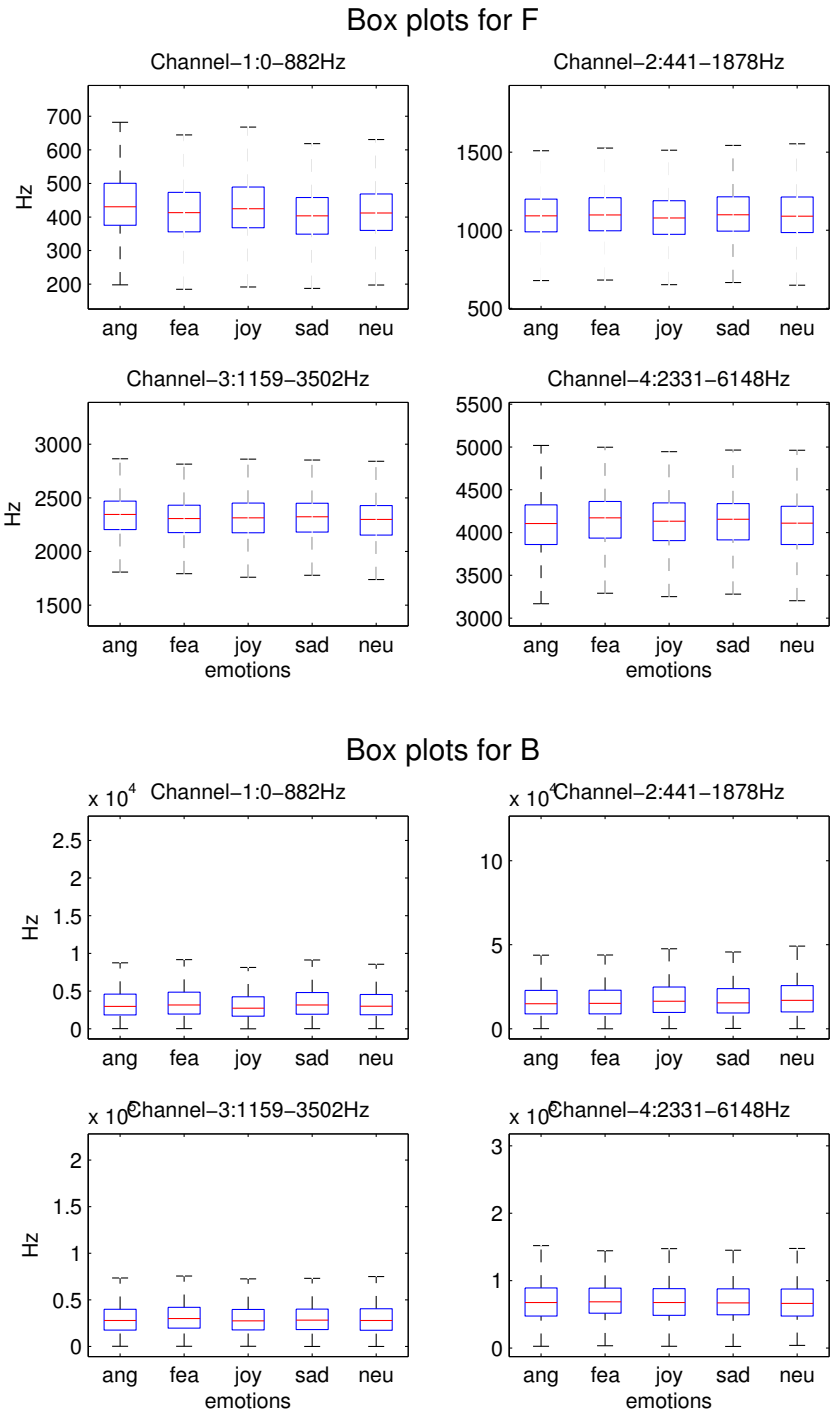
Από τα σχήματα αυτά, μπορούμε να επιβεβαιώσουμε την παρατήρηση που είχαμε κάνει και στη βάση δεδομένων Berlin DB. Ο θυμός και η χαρά παρουσιάζουν παρόμοια μορφολογία μεταξύ τους και αντιδιαστέλλονται με το φόβο και τη λύπη, συναισθήματα που επίσης ορισμένες φορές μοιάζουν μεταξύ τους.



Σχήμα 9.9: Χρονική εξέλιξη του F για μία πρόταση κάθε συναισθήματος στον ομιλητή 18. Οι εκφωνήσεις δεν έχουν την ίδια χρονική διάρκεια για κάθε συναίσθημα



Σχήμα 9.10: Φασματόγραμμα του Ampl mean και του TEO-Auto-Env ορισμένων προτάσεων για τον ομιλητή 18.



Σχήμα 9.11: Box plots των F και B για όλες τις προτάσεις.

9.4 Ταξινόμηση Συναισθημάτων στη Βάση Δεδομένων Αιγινήτειου Νοσοκομείου

Ταξινομούμε τις προτάσεις της βάσης δεδομένων στα 5 συναισθήματα με 2 τρόπους:

- Ανά άτομο: χωρίζουμε τις προτάσεις ανά άτομο και χρησιμοποιούμε μη επιβλεπόμενη ταξινόμηση με GMMs με βάση τα AM-FM χαρακτηριστικά αποδιαμόρφωσης.
- Όλα τα άτομα: χρησιμοποιούμε όλες τις καταγραφόμενες προτάσεις που υπάρχουν στη βάση και εκτελούμε μη επιβλεπόμενη ταξινόμηση με GMMs με βάση τα AM-FM χαρακτηριστικά αποδιαμόρφωσης.

Μη επιβλεπόμενη ταξινόμηση ανά άτομο χρησιμοποιείται γιατί οι προτάσεις είναι σχετικά λίγες ανά άτομο, 35 στο πλήθος, οπότε δεν έχει πρακτική αξία να τις χωρίσουμε σε σύνολο εκπαίδευσης και ελέγχου για επιβλεπόμενη ταξινόμηση.

9.4.1 Ταξινόμηση ανά Άτομο με GMMs

Τα 16 άτομα που συμμετείχαν στις καταγραφές δεν είχαν πάντα μεταξύ τους τον ίδιο τρόπο έκφρασης συναισθήματος. Για το λόγο αυτό επιχειρούμε την ταξινόμηση *ανά άτομο* των συναισθημάτων με GMMs και με βάση τα AM-FM χαρακτηριστικά διαμόρφωσης. Έτσι, μπορούμε να διαπιστώσουμε την επίδραση της προσωπικής έκφρασης κάθε ανθρώπου στη διαμόρφωση του συναισθήματος. Επίσης, αυτός είναι ένας καλός τρόπος να δούμε αν ορισμένα άτομα δεν μπόρεσαν να διαχωρίζουν τα διάφορα συναισθήματα κατά την προφορική έκφραση, οπότε και οι μετρήσεις αυτών μπορούν να παραλειφθούν από τη βάση για να βελτιωθεί η αξιοπιστία της.

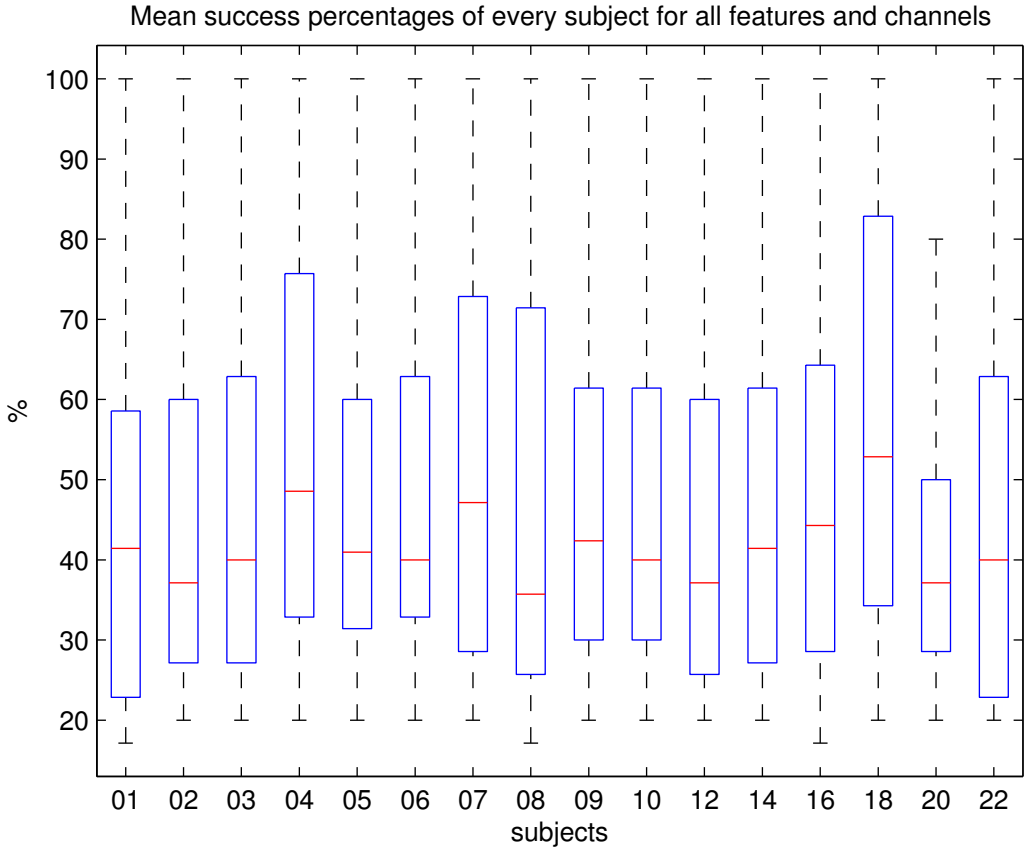
Στους πίνακες 9.2 και 9.3 φαίνονται ενδεικτικά τα αποτελέσματα του πειράματος ταξινόμησης για κάθε χαρακτηριστικό υπολογισμένο σε 12 κανάλια και άτομο. Σημειώνουμε ότι έχουμε παραλείψει τα χαρακτηριστικά για τα οποία η τα ποσοστά επιτυχίας ήταν ίδια με την τυχαία επιλογή (20%) για όλους τους ομιλητές.

Αντίθετα, τα χαρακτηρστικά στιγμιαίας συχνότητας, στην ταξινόμηση ανεξάρτητη του ομιλητή δεν έχουν μεγάλη επιτυχία. Παρόλα αυτά, το ποσοστό των χαρακτηριστικών στιγμιαίας συχνότητας αυξάνεται κατά την ταξινόμηση για κάθε ομιλητή, πράγμα που σημαίνει ότι αυτά επηρεάζονται από τον ομιλητή.

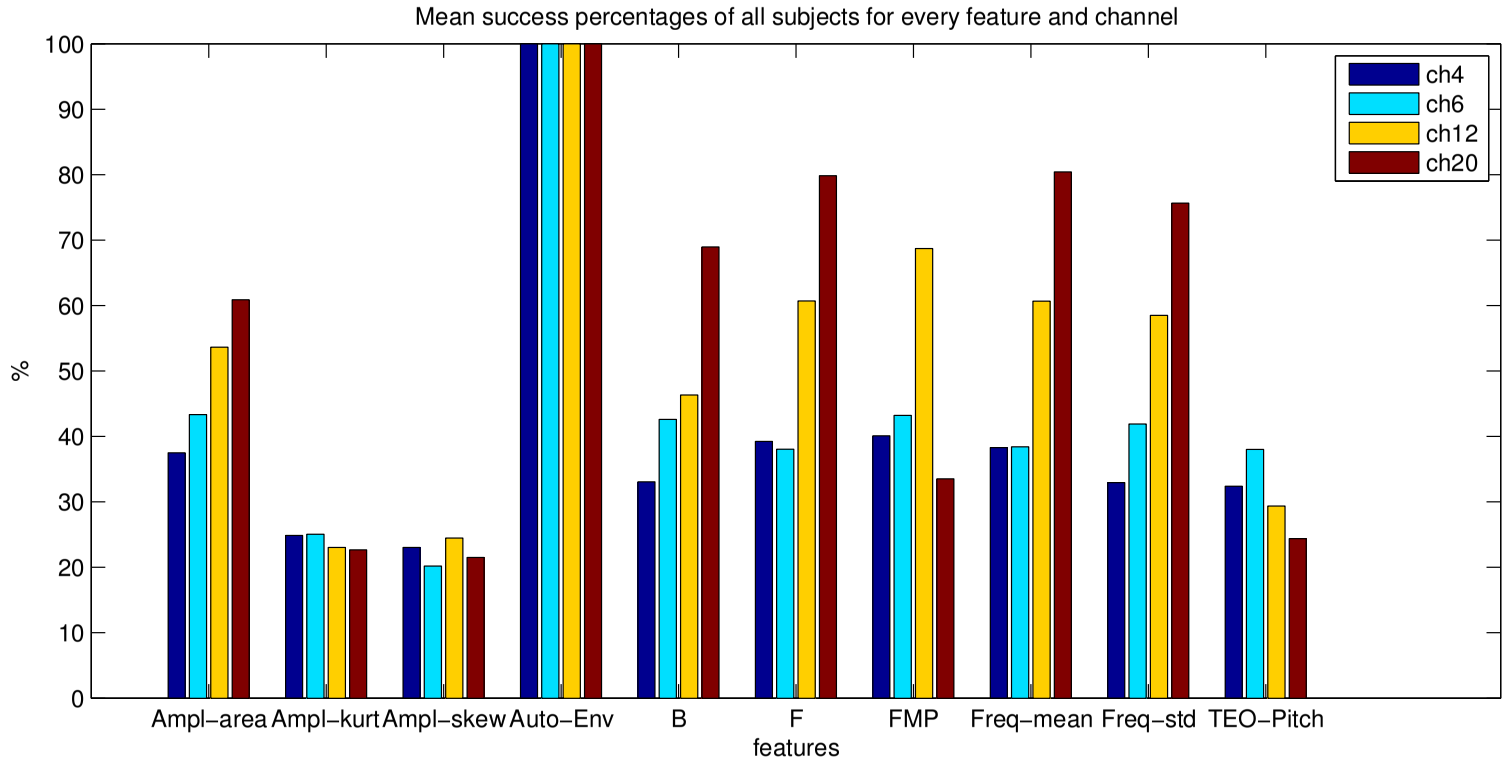
Στο σχήμα 9.12 σχεδιάζουμε τα box plots των ποσοστών επιτυχίας για κάθε άτομο σε όλα τα χαρακτηριστικά και όλα τα κανάλια. Όπως παρατηρούμε, υπάρχουν άτομα, όπως το 04 και το 18, που έχουν υψηλότερα ποσοστά επιτυχίας, σε αντίθεση με άλλα, όπως το 01 και το 20 με μικρά ποσοστά. Αυτό μπορεί να εξηγηθεί από το γεγονός ότι τα χαρακτηριστικά που υπολογίσαμε ταίριαζαν καλύτερα σε κάποια άτομα και σε κάποια άλλα όχι, είτε από το ότι δεν υπήρχε μεγάλη διακριτότητα στο λόγο κάποιων ατόμων ανάμεσα στα συναισθήματα.

Για να δούμε την απόδοση των χαρακτηριστικών σε όλα τα άτομα, σχεδιάζουμε ένα ραβδόγραμμα, που φαίνεται στο σχήμα 9.13. Παρατηρούμε ότι το Auto-Env έχει άριστη απόδοση. Όπως θα δούμε στη συνέχεια, το χαρακτηριστικό αυτό δίνει καλά αποτελέσματα και στην ταξινόμηση ανεξάρτητη του ομιλητή. Υψηλά ποσοστά επιτυχίας για μεμονωμένα άτομα, ιδιαίτερα για 20 κανάλια, αποδίδουν και τα χαρακτηριστικά στιγμιαίας συχνότητας, όπως F, B, Freq mean και Freq std. Αντίθετα, στην επόμενη παράγραφο θα δούμε ότι τα χαρακτηριστικά αυτά δεν είναι ισχυρά για την ταξινόμηση όλων των προτάσεων της

βάσης. Αυτό συμβαίνει γιατί εξαρτώνται από τη μοροφολογία του φωνητικού σωλήνα κάθε ανθρώπου, πράγμα που πιθανώς δεν ισχύει για το Auto-Env.



Σχήμα 9.12: Μέσος όρος ποσοστών επιτυχίας με επιβλεπόμενης ταξινόμησης με GMMs για κάθε άτομο με βάση όλα τα χαρκτηριστικά και όλα τα κανάλια.



Σχήμα 9.13: Μέσος όρος ποσοστών επιτυχίας με επιβλεπόμενης ταξινόμησης με GMMs για κάθε χαρακτηριστικό και κανάλι με βάση όλα τα άτομα.

Χαρ.	**Άτομο**								
	01	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**M.O.**
Ampl-area	57.14	57.14	54.29	71.43	32.86	40	57.14	65.71	**54.46**
Ampl-kurt	20	20	22.86	20	20	34.29	20	28.57	**23.21**
Ampl-skew	20	20	20	37.14	20	37.14	31.43	20	**25.71**
Auto-Env	100	100	100	100	100	100	100	100	**100**
B	51.43	31.43	54.29	45.71	40.95	54.29	54.29	31.43	**45.48**
F	57.14	40	77.14	71.43	50.48	62.86	71.43	77.14	**63.45**
FMP	68.57	68.57	77.14	77.14	49.52	62.86	74.29	54.29	**66.55**
Freq-mean	45.71	51.43	57.14	60	66.67	60	74.29	62.86	**59.76**
Freq-std	60	42.86	51.43	77.14	40.95	68.57	74.29	45.71	**57.62**
TEO-Pitch	31.43	20	45.71	22.86	21.43	37.14	31.43	28.57	**29.82**
M.O.	**51.14**	**45.14**	**56**	**58.29**	**44.29**	**55.71**	**58.86**	**51.43**	

Πίνακας 9.2: Μέσος όρος των ποσοστών επιτυχίας των GMMs για την ταξινόμηση 7 συναισθημάτων σε κάθε άτομο με βάση AM-FM χαρακτηριστικά υπολογισμένα σε 12 κανάλια.

Χαρ.	**Άτομο**								
	09	**10**	**12**	**14**	**16**	**18**	**20**	**22**	**M.O.**
Ampl-area	79.52	40	34.29	54.29	42.86	74.29	42.86	54.29	**52.8**
Ampl-kurt	20	28.57	20	22.86	20	22.86	20	28.57	**22.86**
Ampl-skew	20	34.29	31.43	20	20	20	20	20	**23.21**
Auto-Env	100	100	100	100	100	100	100	100	**100**
B	46.19	42.86	62.86	34.29	48.57	62.86	37.14	42.86	**47.2**
F	43.81	57.14	40	77.14	54.29	88.57	45.71	57.14	**57.98**
FMP	67.14	62.86	80	62.86	74.29	80	71.43	68.57	**70.89**
Freq-mean	55.71	60	60	71.43	57.14	85.71	42.86	60	**61.61**
Freq-std	60.95	68.57	71.43	57.14	68.57	60	48.57	40	**59.4**
TEO-Pitch	40	25.71	25.71	28.57	34.29	34.29	20	22.86	**28.93**
M.O.	**53.33**	**52**	**52.57**	**52.86**	**52**	**62.86**	**44.86**	**49.43**	

Πίνακας 9.3: Μέσος όρος των ποσοστών επιτυχίας των GMMs για την ταξινόμηση 7 συναισθημάτων σε κάθε άτομο με βάση AM-FM χαρακτηριστικά υπολογισμένα σε 12 κανάλια(συνέχεια).

9.4.2 Ταξινόμηση Όλων των Προτάσεων με GMMs

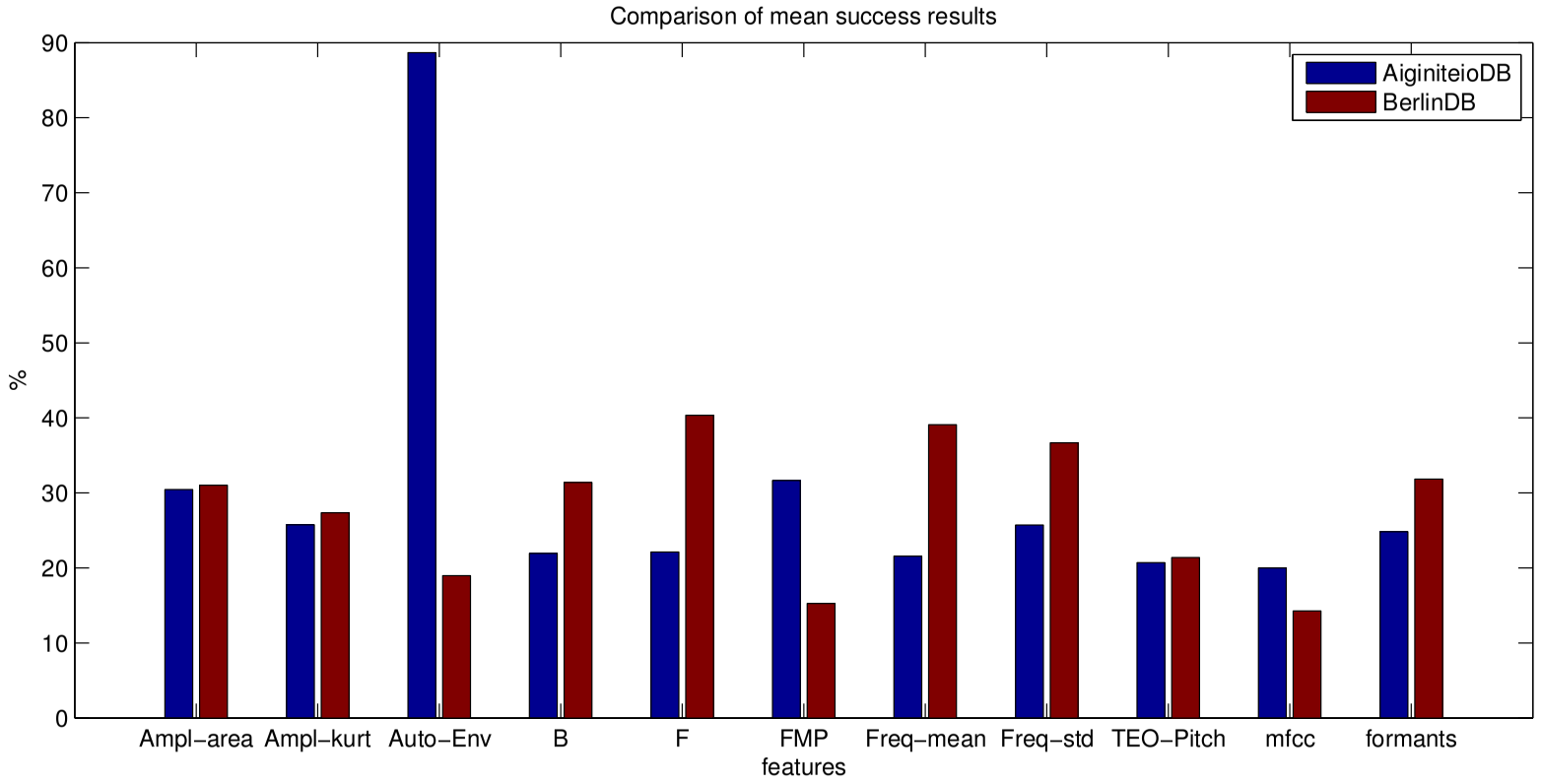
Επιχειρούμε μη επιβλεπόμενη ταξινόμηση με GMMs σε όλη τη βάση δεδομένων του Αιγινήτειου νοσοκομείου. Από τον πίνακα 9.4 προκύπτει ότι το Auto Env χαρακτηριστικό είναι πολύ ισχυρό και δίνει καλά αποτελέσματα ανεξαρτήτως ομιλητή, όπως είδαμε και στην προηγούμενη παράγραφο. Όλα τα υπόλοιπα χαρακτηριστικά δίνουν πιο μέτρια αποτελέσματα, που μπορεί να οφείλονται σε μη επιτυχημένη προσπάθεια έκφρασης του κατάλληλου συναισθήματος από τα άτομα ή και σε αδυναμία των χαρακτηριστικών.

Συγκρίνοντας με τη βάση δεδομένων Berlin DB, όπως φαίνεται στο σχήμα 9.14, μπορούμε να διαπιστώσουμε ότι το Auto-Env δεν είχε τόσο καλά αποτελέσματα σε αυτήν, όσο είχε

στη βάση του Αιγινήτειου Νοσοκομείου. Αντίθετα, τα χαρακτηριστικά συχνότητας παρουσίασαν μεγάλη αδυναμία στη βάση του Αιγινήτειου, που δεν υπήρχε στην Berlin DB. Αυτό μπορεί να δικαιολογηθεί ως εξής: τα άτομα που συμμετείχαν στην έρευνα του Αιγινήτειου νοσοκομείου έδωσαν μεγαλύτερη έμφαση στην ένταση του λόγου για τη διατύπωση των διαφόρων συναισθημάτων και όχι τόσο στη χρειά της ομιλίας. Η ένταση επηρεάζει κατά κόρον το στιγμιαίο πλάτος και άρα και το Auto-Env χαρακτηριστικό, ενώ η χρειά επιδρά στη στιγμιαία συχνότητα και άρα στα Freq-mean, Freq-std, F και B. Πρέπει να λάβουμε υπόψη μας και το γεγονός ότι τα άτομα που συμμετείχαν στη δημιουργία της βάσης δεδομένων του Αιγινήτειου Νοσοκομείου δεν ήταν επαγγελματίες ηθοποιοί, πράγμα που ισχύει για τη βάση Berlin DB, και έτσι δεν μπορούσαν πάντα να χρωματίσουν με την κατάλληλη συναισθηματική επένδυση τις προτάσεις.

	Ομαλοποίηση+LDA			
Χαρακτηριστικό	Αριθμός καναλιών			
	4	6	12	20
Energy-mean	20	20	20	20
Energy-std	20	20	20	20
Ampl-area	30.09	29.01	30.45	27.57
Ampl-kurt	23.06	26.61	25.75	23.59
Ampl-mean	20	20	20	20
Ampl-skew	20	20	20	20
Ampl-std	20	20	20	20
AmplDer1-mean	20	20	20	20
AmplDer1-std	20	20	20	20
AmplDer2-mean	20	20	20	20
AmplDer2-std	20	20	20	20
Auto-Env	97.66	97.66	88.68	93.34
B	21.94	22.66	21.94	29.48
F	23.75	22.86	22.13	33.44
FMP	31.12	27.52	31.65	28.25
Freq-mean	23.94	21.59	21.57	29.82
Freq-std	28.09	27.87	25.71	33.43
TEO-Pitch	20.87	23.06	20.72	23.96
mfcc	20	20	20	20
pitch	26.79			
formants	24.82			

Πίνακας 9.4: Μέσος όρος των ποσοστών επιτυχίας μη επιβλεπόμενης ταξινόμησης με GMMs για τη βάση του Αιγινήτειου Νοσοκομείου.



Σχήμα 9.14: Σύγκριση των ποσοστών επιτυχίας μη επιβλεπόμενης ταξινόμησης με GMMs στις βάσεις δεδομένων Berlin DB και Αιγινίτειου Νοσοκομείου για AM-FM χαρακτηριστικά διαμόρφωσης υπολογισμένα σε 12 κανάλια.

Τέλος, εκτελούμε επιβλεπόμενη ταξινόμηση των προτάσεων της βάσης του Αιγινήτειου Νοσοκομείου. Χρησιμοποιούμε το 70% των προτάσεων για εκπαίδευση και το 30% για έλεγχο. Μελετάμε κάθε χαρακτηριστικό ξεχωριστά σε 4, 6, 12 και 20 κανάλια και εκπαιδεύουμε GMMs με πλήθος γκαουσιανών 1, 2, 4, 8, 16, 24 και 32. Τα αποτελέσματα φαίνονται στους πίνακες 9.5 και 9.6. Όπως και στη μη επιβλεπόμενη ταξινόμηση, βλέπουμε ότι τα καλύτερα ποσοστά επιτυχίας προκύπτουν για το Auto-Env και πιο ισχυρή είναι η ταξινόμηση με 8 γκαουσιανές για χαρακτηριστικα υπολογισμένα σε 12 ή 20 κανάλια.

	1 γκαουσιανή				**2 γκαουσιανές**				**4 γκαουσιανές**			
Χαρακτηριστικό	Αριθμός καναλιών				Αριθμός καναλιών				Αριθμός καναλιών			
	4	6	12	20	4	6	12	20	4	6	12	20
Energy-mean	0.31	0.3	0.27	0.27	0.32	0.3	0.3	0.35	0.33	0.28	0.36	0.3
Energy-std	0.34	0.32	0.33	0.27	0.36	0.3	0.32	0.29	0.35	0.28	0.28	0.3
Ampl-area	0.29	0.29	0.29	0.28	0.35	0.3	0.27	0.29	0.33	0.28	0.31	0.33
Ampl-kurt	0.21	0.16	0.19	0.19	0.27	0.23	0.21	0.22	0.21	0.25	0.19	0.21
Ampl-mean	0.26	0.33	0.3	0.3	0.28	0.36	0.32	0.37	0.26	0.31	0.32	0.37
Ampl-skew	0.26	0.27	0.21	0.19	0.29	0.27	0.3	0.25	0.29	0.27	0.22	0.21
Ampl-std	0.27	0.3	0.3	0.31	0.27	0.29	0.32	0.39	0.27	0.3	0.37	0.32
AmplDer1-mean	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19
AmplDer1-std	0.26	0.28	0.35	0.32	0.26	0.28	0.32	0.32	0.26	0.28	0.3	0.31
AmplDer2-mean	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19
AmplDer2-std	0.26	0.26	0.24	0.3	0.27	0.28	0.36	0.37	0.27	0.28	0.34	0.3
Auto-Env	0.61	0.62	0.51	0.65	0.59	0.61	0.58	0.67	0.59	0.64	0.61	0.55
B	0.2	0.33	0.22	0.3	0.29	0.3	0.2	0.28	0.26	0.27	0.27	0.33
F	0.35	0.32	0.36	0.28	0.3	0.29	0.39	0.33	0.29	0.32	0.35	0.33
Freq-mean	0.3	0.3	0.33	0.29	0.31	0.31	0.36	0.28	0.28	0.28	0.27	0.36
Freq-std	0.23	0.26	0.22	0.31	0.22	0.32	0.25	0.34	0.22	0.29	0.27	0.28
TEO-Pitch	0.22	0.24	0.2	0.22	0.24	0.2	0.19	0.2	0.19	0.21	0.18	0.19
mfcc	0.29	0.3	0.31	0.31	0.29	0.3	0.29	0.32	0.28	0.27	0.3	0.31
pitch2	0.31				0.32				0.24			
formants	0.25				0.28				0.34			
	0.28	**0.29**	**0.28**	**0.29**	**0.29**	**0.3**	**0.3**	**0.31**	**0.28**	**0.29**	**0.29**	**0.3**

Πίνακας 9.5: Μέσος όρος των ποσοστών επιτυχίας των GMMs με *1, 2 και 4 γκαουσιανές ανά ομάδα* για την ταξινόμηση 5 συναισθημάτων.

	8 γκαουσιανές				**16 γκαουσιανές**				**24 γκαουσιανές**				**32 γκαουσιανές**			
Χαρακτηριστικό	Αριθμός καναλιών				Αριθμός καναλιών				Αριθμός καναλιών				Αριθμός καναλιών			
	4	6	12	20	4	6	12	20	4	6	12	20	4	6	12	20
Energy-mean	0.31	0.3	0.3	0.33	0.3	0.32	0.34	0.3	0.31	0.3	0.33	0.32	0.33	0.33	0.35	0.33
Energy-std	0.35	0.33	0.29	0.31	0.35	0.31	0.28	0.31	0.33	0.33	0.31	0.35	0.35	0.35	0.3	0.33
Ampl-area	0.3	0.29	0.34	0.36	0.28	0.36	0.38	0.3	0.32	0.36	0.29	0.3	0.38	0.37	0.35	0.3
Ampl-kurt	0.19	0.19	0.21	0.24	0.24	0.25	0.2	0.24	0.19	0.25	0.21	0.27	0.22	0.24	0.21	0.26
Ampl-mean	0.24	0.35	0.35	0.41	0.25	0.32	0.37	0.35	0.25	0.33	0.36	0.35	0.24	0.33	0.31	0.33
Ampl-skew	0.27	0.22	0.29	0.21	0.28	0.27	0.29	0.21	0.29	0.27	0.24	0.16	0.26	0.25	0.29	0.2
Ampl-std	0.28	0.28	0.36	0.34	0.28	0.26	0.34	0.36	0.27	0.27	0.3	0.36	0.25	0.28	0.36	0.39
AmplDer1-mean	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19
AmplDer1-std	0.26	0.28	0.32	0.32	0.26	0.28	0.32	0.32	0.26	0.28	0.33	0.32	0.26	0.28	0.32	0.32
AmplDer2-mean	0.19	0.09	0.19	0.19	0.19	0.09	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.09	0.19	0.19
AmplDer2-std	0.27	0.28	0.32	0.36	0.27	0.27	0.32	0.38	0.27	0.27	0.32	0.33	0.27	0.28	0.32	0.35
Auto-Env	0.6	0.57	0.59	0.61	0.56	0.51	0.5	0.38	0.64	0.67	0.28	0.36	0.47	0.61	0.27	0.2
B	0.22	0.3	0.27	0.28	0.27	0.24	0.27	0.26	0.22	0.25	0.3	0.26	0.22	0.28	0.25	0.3
F	0.35	0.31	0.32	0.29	0.34	0.33	0.32	0.35	0.34	0.35	0.33	0.27	0.36	0.33	0.33	0.33
Freq-mean	0.34	0.29	0.35	0.35	0.37	0.27	0.36	0.27	0.34	0.29	0.3	0.23	0.36	0.29	0.3	0.28
Freq-std	0.25	0.3	0.25	0.27	0.23	0.27	0.26	0.3	0.25	0.27	0.25	0.27	0.25	0.27	0.25	0.28
TEO-Pitch	0.21	0.2	0.19	0.19	0.22	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19
mfcc	0.28	0.29	0.3	0.3	0.3	0.29	0.32	0.32	0.29	0.28	0.31	0.32	0.28	0.3	0.32	0.3
pitch2	0.34				0.23				0.19				0.19			
formants	0.39				0.35				0.36				0.35			
	0.29	**0.29**	**0.31**	**0.31**	**0.29**	**0.28**	**0.3**	**0.29**	**0.29**	**0.3**	**0.28**	**0.28**	**0.28**	**0.29**	**0.28**	**0.28**

Πίνακας 9.6: Μέσος όρος των ποσοστών επιτυχίας των GMMs με *8, 16, 24 και 32 γκαουσιανές ανά ομάδα* για την ταξινόμηση 7 συναισθημάτων.

Κεφάλαιο 10

Συμπεράσματα και Πιθανές Μελλοντικές Επεκτάσεις

Το πεδίο έρευνας της αναγνώρισης συναισθήματος μέσα από τη φωνή είναι ευρύ και μπορεί να μελετηθεί ποικιλλοτρόπως. Στη διπλωματική εργασία δώσαμε έμφαση στην εξαγωγή κατάλληλων χαρακτηριστικών, που μπορούν να εντοπίσουν τις συναισθηματικό χρωματισμό της φωνής. Στη συνέχεια, εξετάσαμε τα χαρακτηριστικά αυτά σε 2 βάσεις δεδομένων: τη γερμανική Berlin Database of emotional speech και την ελληνική βάση του Αιγινήτειου Νοσοκομείου.

Έγινε μία επισκόπηση των βασικών χαρακτηριστικών που χρησιμοποιήθηκαν σε προηγούμενες έρευνες, όπως το pitch, τα formants και η διάρκεια εκφοράς και παρατηρήθηκαν διαφορές μεταξύ ορισμένων συναισθημάτων. Βρέθηκε ότι ο θυμός και η χαρά έχουν υψηλές τιμές pitch, ενώ η πλήξη, η λύπη και το ουδέτερο πιο μικρές τιμές. Επίσης, η απαρέσκεια και η λύπη εκφράζονται με πιο αργό λόγο, ενώ η χαρά, ο φόβος και το ουδέτερο και πιο κοφτό λόγο.

Όσον αφορά στα χαρακτηριστικά της φωνής, διαπιστώσαμε ότι τα AM-FM χαρακτηριστικά διαμόρφωσης μπορούν να εντοπίσουν με αρκετή ακρίβεια το συναίσθημα στο λόγο. Μάλιστα, στα περισσότερα πειράματα, η απόδοσή τους ξεπερνάει τα κλασσικά χαρακτηριστικά του pitch, των formants και τα mfcc. Τα χαρακτηριστικά στιγμιαίου πλάτους με μεγάλη επιτυχία αναγνώρισης είναι το εμβαδόν στιγμιαίου πλάτους σε κάθε παράθυρο (Ampl-area) και το εμβαδόν της αυτοσυσχέτισης του πλάτους (Auto-Env). Υπάρχουν περιπτώσεις που το ένα υπερτερεί του άλλου και αντίστροφα, για αυτό και δεν μπορούμε να πούμε με σιγουριά πιο είναι το ισχυρότερο. Οι στατιστικές ροπές του στιγμιαίου πλάτους και κυρίως ο μέσος όρος και η κύρτωση επιφέρουν επίσης σχετικά καλά αποτελέσματα. Από τα χαρακτηριστικά στιγμιαίας συχνότητας, ο απλός και σταθμισμένος όρος στιγμιαίας συχνότητας δίνουν υψηλά ποσοστά επιτυχίας σχεδόν σε κάθε ταξινομητή. Τα χαρακτηριστικά αυτά εντοπίζουν τις διαμορφώτριες συχνότητες του σήματος φωνής με ακρίβεια και για αυτό το λόγο είναι πολύ αποδοτικά.

Οι μέθοδοι ταξινόμησης που δοκιμάσαμε είναι ο K-means τα GMMs και τα αναδιαμορφούμενα GMMs. Στην ταξινόμηση με K-means βρέθηκε ότι υπερτερούν τα ομαλοποιημένα χαρακτηριστικά διαμόρφωσης υπολογισμένα σε 20 κανάλια. Η ομαλοποίηση είναι πιθανό να απομακρύνει την περιττή πληροφορία και να διατηρεί μόνο την πληροφορία που είναι απαραίτητη για την αναγνώριση του συναισθήματος. Στα GMMs ο πιο αποδοτικός αριθμός γκαουσιανών ανά κλάση παρατηρήθηκε ότι είναι το 24 και τα πιο ισχυρά χαρακτηριστικά εί-

ναι αυτά που έχουν υπολογιστεί σε 12 κανάλια. Πιστεύουμε ότι ο αριθμός των 12 καναλιών είναι ο ιδανικός ώστε να διατηρείται η χρήσιμη πληροφορία και να μην υπεισέρχεται μεγάλη λεπτομέρεια στους υπολογισμούς. Τέλος, η ταξινόμηση με δυναμικά διαμορφούμενο μείγμα γκαουσιανών, αν και μη επιβλεπόμενη, επιφέρει πολύ καλά αποτελέσματα. Ο λόγος είναι ότι σε αυτά το πλήθος των ομάδων καθορίζεται δυναμικά. Δεν είναι απαραίτητο το πλήθος των κλάσεων να ισούται με το πλήθος των συναισθημάτων υπό ταξινόμηση, οπότε διανύσματα που αφορούν ένα συναίσθημα μπορεί να ανήκουν σε πολλές ομάδες.

Παρόλα αυτά δεν έχουμε χρησιμοποιήσει όλη τη δύναμη των ταξινομητών της αναγνώρισης προτύπων. Ένα μεγάλο κεφάλαιο που δεν έχει εξεταστεί είναι ο έλεγχος των χαρακτηριστικών που μελετήσαμε με HMMs, ώστε να ληφθεί υπόψη και ο παράγοντας του χρόνου. Επίσης, όσον αφορά στα χαρακτηριστικά ταξινόμησης, ενδιαφέρον θα ήταν να βρεθούν κατάλληλα σταθμισμένοι συνδυασμοί των χαρακτηριστικών διαμόρφωσης AM-FM που πιθανώς αυξάνουν και τα ποσοστά επιτυχούς ταξινόμησης. Τέλος, ένα χρήσιμο πείραμα για τη βάση του Αιγινήτειου Νοσοκομείου θα ήταν να ειπωθούν οι ίδιες προτάσεις από επαγγελματίες ηθοποιούς, ώστε να διαπιστωθεί σε ποια χαρακτηριστικά διαφέρουν από τους απλούς ανθρώπους.

Γενικά, οι τεχνικές αναγνώρισης συναισθήματος είτε μέσω ήχου είτε μέσω εικόνας είναι ένας μεγάλος κλάδος με μεγάλη ανάπτυξη τα τελευταία χρόνια, που ανήκει στο πεδίο έρευνας του affective computing. Πλέον δεν αρκεί ένας υπολογιστής να κάνει μόνο υπολογισμούς και εργασίες, αλλά αναζητούνται τρόποι να γίνει πιο φιλικός και επικοινωνιακός προς το χρήστη. Όπως υποστηρίζει άλλωστε και μία ομάδα ερευνητών [62], προφανώς, δεν είναι απαραίτητο οι μηχανές να αναπτύξουν τη συναισθηματική ικανότητα των ανθρώπων, αλλά είναι απαραίτητο να έχουν μία συναισθηματική νοημοσύνη. Αν για παράδειγμα μία μηχανή μας μιλάει, αλλά δε μας ακούει ποτέ, τότε είναι πολύ πιθανό να ενοχληθούμε, με παρόμοιο τρόπο που θα μας ενοχλούσε αν ένας άνθρωπος μας μιλούσε αλλά δε μας άκουγε.

Bibliography

[1] S. Abrilian, L. Devillers, and J.-C. Martin. Emotv1: Annotation of real-life emotions for the specification of multimodal affective interfaces. *HCI International, Las Vegas, July*, 2005.

[2] L.F. Barrett and E. Bliss-Moreau. Affect as a psychological primitive. *Advances in Experimental Social Psychology, Academic Press*, 41:167–193, 2009.

[3] M. E. Beckman. http://www.ling.ohio-state.edu/ tobi/. 1999.

[4] C.M. Bishop. *Pattern Recogniiton and Machine Learning*. Springer, 2006.

[5] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio if a sampled sound. *IEA Proc.*, 17, 1993.

[6] M. Bulut, S. Lee, and S. Narayanan. Recognition for synthesis: Automatic parameter selection for resynthesis of emotional speech from neutral speech. *Proceedings of ICASSP, Las Vegas, Nevada, April 2008*.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss. A database of german emotional speech. *Proceedings of Interspeech 2005, Lissabon, Portugal*, 2005.

[8] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 2008.

[9] C. Busso, S. Lee, and S. Narayanan. Using neutral speech models for emotional speech analysis. *Proceedings of InterSpeech ICSLP, Antwerp, Belgium, August 2007*.

[10] C. Busso and S. Narayanan. Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the iemocap database. *Interspeech 2008 - Eurospeech, Brisbane, Australia, September 2008*.

[11] D.A. Cairns and J.H.L. Hansen. Nonlinear analysis and classification of speech under stressed conditions. *J. Acoust. Soc. Am.*, 96:3392–3400, 1994.

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

[13] J. Cichosz and K. Slot. Emotion recognition in speech signal using emotion-extracting binary decision trees. *Affective Computing and Intelligent Interfaces (ACII)*, 2007.

[14] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. Mcmahon, M. Sawey, and M. Schroder. 'feeltrace': An instrument for recording perceived emotion in real time. *In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

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

[16] C. Cullen, B. Vaughan, S. Kousidis, and J. McAllen. Emotional speech corpus construction, annotation and distribution. *Proceedings of sixth international conference on Language Resources and Evaluation (LREC)*, 2008.

[17] C. Cullen, B. Vaughan, S. Kousidis, and Y. Wang. Generation of high quality audio natural emotional speech corpus using task based mood induction. *InSciT, Merida*, 2006.

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

[19] L. Devillers, R. Cowie, J-C. Martin, E. Douglas-Cowie, S. Abrilian, and M. McRorie. Real life emotions in french and english tv video clips: an integrated annotation protocol combining continuous and discrete approaches. *International Conference on Language Resources and Evaluation*, page 1105P1110, 2006.

[20] D. Dimitriadis and P. Maragos. Continuous energy demodulation methods and application to speech analysis. *Elsevier Speech Communication*, 48:819–837, 2005.

[21] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach. Emotional speech: Towards a new generation of databases. *Elsevier Speech Communication*, 40:33P60, 2003.

[22] E. Douglas-Cowie, L. Devillers, J. C. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox. Multimodal databases of everyday emotion: Facing up to complexity. *Proc. Interspeech 2005, Lisbon, Portugal*, page 813P816, 2005.

[23] P. Ekman. Facial expressions of emotion. *American Psychologist*, 48:384–392, 1993.

[24] P. Ekman. Basic emotions. *Handbook of Cognition and Emotion, John Wiley Sons*, pages 45–60, 1999.

[25] M. Faundez-Zanuy, S. McLaughlin, A Esposito, A. Hussain, J. Schoentgen, G. Kubin, W.B. Kleijn, and P. Maragos. Nonlinear speech processing: Overview and applications. *Control and Intelligent Systems*, 30:1–10, 2002.

[26] F.Dellaert. Matlab clustering package, version 2. 2003.

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

[28] R. Fernandez and R.W. Picard. Modeling drivers' speech under stress. *Speech Communication*, 40:145–159, 2003.

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

[30] A. Gerrards-Hesse, K. Spies, and F.W. Hesse. Experimental inductions of emotional states and their effectiveness: A review. *British Journal of Psychology*, 85:55–78, 1994.

[31] H.J. Go, K.C. Kwak, D.J. Lee, and M.G. Chun. Emotion recognition from the facial image and speech signal. *SICE Annual Conference in Frankfurt*, 2003.

[32] S. Greenberg and B.E.D. Kingsbury. The modulation spectrogram: In pursuit of an invariant representation of speech. *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'97)*, 19:1647–1650, 1997.

[33] J.H. Hansen, L. Gavidia-Ceballos, and J.F. Kaiser. A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment. *IEEE Trans Biomed Eng.*, 3:300–313, 1998.

[34] J.H.L. Hansen, S.E. Bou-Ghazale, R. Sarikaya, and B. Pellom. Getting started with the susas: Speech under simulated and actual stress database. *Robust Speech ProcessingLaboratory, Technical Report*, 1998.

[35] J.H.L. Hansen and S.A. Patil. Speech under stress: Analysis, modeling and recognition. *Speaker Classification I: Fundamentals, Features, and Methods. Ed. Christian Muller. Berlin, Heidelberg: Springer-Verlag*, pages 108–137, 2007.

[36] J.H.L. Hansen and B.D. Womack. Feature analysis and neural network-based classification of speech under stress. *IEEE Trans. Speech Audio Process*, 4:307–313, 1996.

[37] I.A. Iliev and M.S. Scordilis. Emotion recognition in speech using inter-sentence glottal statistics. *Proceedings of the 15th International Conference on systems, Signals and Image Processing (IWSSIP 2008), Bratislava, Slovakia, June 2008*, pages 465–468, 2008.

[38] I.A. Iliev, Y. Zhang, and M.S. Scordilis. Spoken emotion classification using tobi features and gmms. *IEEE 6th EURASIP Conference focused on Speech and Image Processing*, pages 495–498, 2007.

[39] I.R.Murray and J.L.Arnott. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America (JASA)*, 93:1097–1108, 1993.

[40] D.N. Jiang and L.H. Cai. Speech emotion classification with the combination of statistic features and temporal features. *IEEE International Conference on Multimedia and Expo*, 3:1967–1970, 2004.

[41] J.F. Kaiser. On a simple algorithm to calculate the 'energy' of a signal. *IEEE Int. Conf. Acoust. Speech, Signal Processing, Albuquerque, NM, April 1990*, pages 381–384, 1990.

[42] A. B. Kandali, A. Routray, and T.K. Basu. Emotion recognition from assamese speeches using mfcc features and gmm classifier. *Proc. IEEE region 10 conference TENCON 2008, 19P21 Nov., Hyderabad, India*, pages 1–5, 2008.

[43] A. B. Kandali, A. Routray, and T.K. Basu. Emotion recognition from speeches of some native languages of assam independent of text and speaker. *National Seminar on Devices, Circuits and Communication, Department of E.C.E., B.I.T. Mesra, Ranchi, Jharkhand, India*, 2008.

[44] J. Kim, S. Lee, and S. Narayanan. An explanatory study of manifolds of emotional speech. *Proceedings of ICASSP, Dallas*, 2010.

[45] S. Kim, P. G. Georgiou, S. Lee, and S. Narayanan. Real-time emotion detection system using speech: Multi-modal fusion of different timescale features. In *Proc. of IEEE Multimedia Signal Processing Workshop, Chania, Greece*, 2007.

[46] O-W Kwon, K Chan, J Hao, and T-W Lee. Emotion recognition by speech signals. *Proc. of Eurospeech. 2003, Genewa*, pages 125–128, 2003.

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

[48] 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, October 2004*.

[49] C.L. Lisetti and F. Nasoz. Affective intelligent car interfaces with emotion recognition. *11th International Conference on Human Computer Interaction (HCI), July 22-27, Las Vegas, USA*, 2005.

[50] R.W. Schafer L.R. Rabiner. *Digital Processing of Speech Signals*. Prentice-Hall Signal Processing Series, 1978.

[51] P. Maragos, J.F Kaiser, and T.F. Quatieri. Energy separation in signal modulations with application to speech analysis. *IEEE Trans. Signal Proc.*, 41:3025–3051, 1993.

[52] E. Moore, M.A Clements, J.W. 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*, 55:96–107, 2008.

[53] D. Morrison, R. Wang, and L. C. De Silva. Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, 49:98–112, 2007.

[54] S.J.L. Mozziconacci and D.J. Hermes. Expression of emotion and attitude through temporal speech variations. *Proc. 2000 Int. Conf. Spoken Language Processing (ICSLP)*, pages 1–24, 2000.

[55] A. Mpoutri. Organization and structure of emotions (in greek). `http://www.positiveemotions.gr/index.php?option=com_content&task=view&id=28&Itemid=55`, September 2005.

[56] R. Nakatsu, J. Nicholson, and N. Tosa. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. In *IEEE 3rd Workshop on Multimedia Signal Processing*, pages 439 – 444, 1999.

[57] T.L. Nwe, S.W. Foo, and L.C. De Silva. Speech emotion recognition using hidden markov models. *Elsevier Speech Communication*, 41:603–623, 2003.

[58] A. Ortony and T.J Turner. What's basic about basic emotions? *PsychologicalReview*, 97:315–331, 1990.

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

[60] V. Petrushin. Emotion in speech: Recognition and application to call centers. *Artificial Neural Networks in Engineering (ANNIE '99), 7-10 November 1999, St. Louis*, 1999.

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

[62] R.W. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:1175–1191, 2001.

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

[64] R. Plutchik. The nature of emotions. *American Scientist*, 89:344–350, 2001.

[65] A. Potamianos and P. Maragos. Speech formant frequency and bandwidth tracking using multiband energy demodulation. *Journal of Acoustical Society of America*, 99:3795–3806, 1996.

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

[67] J.A. Russell and L. Feldman Barrett. Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology, American Psychological Association*, 76:805–819, 1999.

[68] S.A.Patil and J.H.L Hansen. Detection of speech under physical stress: model development, sensor selection, and feature fusion. *IEEE trans. on SAP*, 9:201–216, 2001.

[69] Klaus R. Scherer. A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology. *Implications for SpeechS, Proceedings of the ICSLP*, pages 379–382, 2000.

[70] K.R. Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44:695–729, 2005.

[71] K.R. Scherer and Semsler. What are emotions? and hotw can they be measured? *Social Science Information*, 44:695–729, 2005.

[72] B. Schuller, S. Reiter, R. Mueller, M. Al-Hames, and G. Rigoll. Speaker-independent speech emotion recognition by ensemble classification. *Proc. ICME 2005, Amsterdam, Netherlands, 2005.*

[73] B. Schuller, S. Reiter, and G. Rigoll. Evolutionary feature generation in speech emotion recognition. *IEEE International Conference on Multimedia and Expo*, pages 5–8, 2006.

[74] B. Schuller, G. Rigoll, and M. Lang. Hidden markov model-based speech emotion recognition. *Proc. of the 2003 International Conference on Multimedia and Expo*, 2:401–404, 2003.

[75] B. Schuller, G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture. *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 5:577–580, 2004.

[76] M. Sedaaghi, C. Kotropoulos, and D. Ververidis. Using adaptive genetic algorithms to improve speech emotion recognition. *Proc. IEEE Workshop Multimedia Signal Processing (MMSP)*, 2007.

[77] J. Sidorova. Speech emotion recognition with tgi+.2 classifier. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 54–60, 2009.

[78] S. Krishnan T. Tabatabaei and A. Guergachi. Speech-based emotion recognition using sequence discriminant support vector machines. *Proc. Canadian Medical and Biological Engineering Conference (CMBEC), Toronto, Ontario*, 2007.

[79] H. Teager and S. Teager. Evidence for nonlinear production mechanisms in the vocal tract. *Speech Production and Speech Modeling, NATO Advanced Study Institute*, 55:241–261, 1990.

[80] J.B. Tenenbaum, V. de Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *SCIENCE*, 290:2319–2323, 2000.

[81] Wikipedia the free encyclopedia. List of emotions. `http://en.wikipedia.org/wiki/List_of_emotions`, June 2010.

[82] N. Tosa and R. Nakatsu. Life-like communication agent - emotion-sensing character "mic" and feeling session character "muse". In *ICMCS*, pages 12–19, 1996.

[83] D. Tsujinishi, Y. Koshiba, and S. Abe. Why pairwise is better than one-against-all or all-at-once. *Proc. if IEEE International Conference on Neural Networks*, 1:693–698, 2004.

[84] D. Ververidis. Digital signal processing techniques for emotion recognition (in greek). *PhD*, 2008.

[85] D. Ververidis and C. Kotropoulos. A state of the art review on emotional speech databases. *Proc. 1st Richmedia Conference, Laussane*, pages 109–119, 2003.

[86] D. Ververidis and C. Kotropoulos. Automatic emotional speech classification. *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'04)*, 1:593–596, 2004.

[87] D. Ververidis and C. Kotropoulos. Automatic speech classification to five emotional states based on gender information. *Proc. European Signal Processing Conf. (EUSIPCO'04)*, 1:341–344, 2004.

[88] D. Ververidis and C. Kotropoulos. Emotional speech classification using gaussian mixture models. *Proc. IEEE Inter. Symposium on Circuits and Systems (ISCAS)*, 2005.

[89] D. Ververidis and C. Kotropoulos. Emotional speech classification using gaussian mixture models and the sequential floating forward selection algorithm. *Proc. Int. Conf. Multimedia and Expo (ICME04)*, 2005.

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

[91] D. Ververidis and C. Kotropoulos. Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition. *Elsevier Signal Processing*, 12:2956–2970, 2008.

[92] D. Ververidis and C. Kotropoulos. Gaussian mixture modeling by exploiting the mahalanobis distance. *IEEE Trans. Signal Processing*, 56:2797–2811, 2008.

[93] Y. Wang and L. Guan. An investigation of speech- based human emotion recognition. *IEEE 6th Workshop on Mul- timedia Signal Processing*, 2004.

[94] A. Wichmann. The attitudinal effects of prosody, and how they relate to emotion. *SpeechEmotion*, pages 143–148, 2000.

[95] B.D. Womack and J.H.L. Hansen. Classification of speech under stress using target driven features. *Elsevier Speech Communication*, 20:131–150, 1996.

[96] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan. An acoustic study of emotions expressed in speech. *Proceedings of ICSLP, Jeju, Korea, October 2004*.

[97] C. Zhou, Y. Zhao, L. Zhao, W. Zhen, and Y. Bao. Emotional recognition using a compensation transformation in speech signal. *Computational Linguistics and Chinese Language Processing*, 12:79–90, 2007.

[98] G. Zhou, J.H.L Hansen, and J.F Kaiser. Classification of speech under stress based on features derived form the nonlinear teager energy operator. *Proc. Int. Conf. Acoustic, Speech, Signal Processing, Seattle, WA, May 12-15*, pages 549–552, 1998.

[99] G. Zhou, J.H.L Hansen, and J.F Kaiser. Nonlinear feature based classification of speech under stress. *IEEE trans. on SAP*, 9:201–216, 2001.