ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Προσαρμογή του Ομιλητή για Αναγνώριση Συναισθήματος από Φωνή

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ευαγγελία Χατζηαγάπη

Επιβλέπων: Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2017

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Προσαρμογή του Ομιλητή για Αναγνώριση Συναισθήματος από Φωνή

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ευαγγελία Χατζηαγάπη

Επιβλέπων: Αλέξανδρος Ποταμιάνος
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή επιτροπή την 6η Οκτωβρίου 2017:

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Αλέξανδρος Ποταμιάνος	**Πέτρος Μαραγκός**	**Κωνσταντίνος Τζαφέστας**
Αναπληρωτής Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2017

(Υπογραφή)

.....
Ευαγγελία Χατζηαγάπη
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π

Περίληψη

Η ένταξη της έννοιας του συναισθήματος στην αλληλεπίδραση ανθρώπου-μηχανής γίνεται όλο και πιο δημοφιλής τα τελευταία χρόνια. Άλλωστε, το συναίσθημα αποτελεί απαραίτητο στοιχείο της ανθρώπινης επικοινωνίας, καθορίζοντας σε σημαντικό βαθμό την αντίληψη της μεταδιδόμενης πληροφορίας από την άλλη πλευρά. Η παρατήρηση αυτή οδηγεί στην ανάγκη έρευνας των ιδιαίτερων ιδιοτήτων του ανθρώπου, που σχετίζονται με τη ψυχολογία του και τις νοητικές του διεργασίες. Σκοπός είναι η ανάπτυξη συστημάτων, ικανών να αντιληφθούν και πιθανώς να προσομοιώσουν ανθρώπινες συναισθηματικές αντιδράσεις.

Εξάγοντας ακουστικά χαρακτηριστικά από σήμα φωνής, επιθυμείται η κατηγοριοποίησή του σε μία συναισθηματική κλάση. Πλήθος μοντέλων Αναγνώρισης Συναισθήματος από Φωνή έχει προταθεί για την υλοποίηση της κατηγοριοποίησης αυτής. Στην πράξη, κάθε μοντέλο από αυτά καλείται να αναγνωρίσει το συναίσθημα μιας εκφώνησης, η οποία πιθανότατα θα προέρχεται από διαφορετικό ομιλητή ή περιβάλλον ηχογράφησης, συγκριτικά με τα δεδομένα εκπαίδευσης. Όμως, η διαφοροποίηση αυτή, συχνά, δε λαμβάνεται υπόψη κατά την ανάπτυξη τέτοιων μοντέλων. Οπότε, το ερώτημα είναι: Πόσο σημαντικά είναι τα διαφορετικά στοιχεία φωνής των ομιλητών, κατά την αναγνώριση συναισθήματος;

Στοιχεία που διαφοροποιούν τη φωνή κάθε ομιλητή από τους υπόλοιπους μπορεί να είναι βιολογικά, όπως το φύλο και η ηλικία, ή κοινωνικο-πολιτισμικά, όπως η γλώσσα, η κουλτούρα και ο προσωπικός χαρακτήρας. Μάλιστα, λόγω της ιδιαίτερης φύσης του συναισθήματος, η έκφρασή του ποικίλει σε σημαντικό βαθμό, ανάλογα με τα παραπάνω στοιχεία. Μια πρώτη προσέγγιση για τη μείωση των διαφορών αυτών είναι η κανονικοποίηση των δεδομένων, με χρήση απλών τεχνικών. Ιδιαίτερο ενδιαφέρον παρουσιάζουν μια σειρά από τεχνικές Προσαρμογής του Ομιλητή, που έχουν αναπτυχθεί στον τομέα της Αυτόματης Αναγνώρισης Φωνής. Στη συγκεκριμένη εργασία, έγινε εφαρμογή και σύγκριση των πιο βασικών από αυτές, με σκοπό την αναγνώριση συναισθήματος. Επιπλέον, ερευνήθηκαν και παραλλαγές τους.

Μια άλλη προσέγγιση, δοθείσας της διαφορετικής έκφρασης των ομιλητών, είναι η εύρεση της ουδέτερης ομιλίας τους. Η γνώση των ουδέτερων χαρακτηριστικών κάθε ομιλητή καθιστά εφικτή την ανίχνευση οποιασδήποτε συναισθηματικής του φόρτισης, με σκοπό τη βελτίωση της αλληλεπίδρασης ανθρώπου-μηχανής. Σε αυτό το πνεύμα, εξελίχθηκε μια ήδη ανεπτυγμένη ιδέα συστήματος με βάση τη βιβλιογραφία. Σημαντικό χαρακτηριστικό του νέου συστήματος αποτελεί η ένταξη τεχνικής Προσαρμογής του Ομιλητή.

Λέξεις κλειδιά: Προσαρμογή Ομιλητή, Αναγνώριση Συναισθήματος, Φωνή

Abstract

The integration of the concept of emotion into the human-computer interaction has become more and more popular in recent years. After all, emotion is an essential element of human communication, defining to a great extent the other side's perception of the transmitted information. This observation leads to the need to investigate the particular qualities of the human, related to his psychology and mental processes. The aim is to develop systems that are capable of understanding and possibly simulating human emotional reactions.

By extracting audio features from a voice signal, it is desired to categorize it in an emotional class. A great number of Speech Emotion Recognition models have been suggested to implement this categorization. In reality, each model is required to recognize the emotion of an utterance, which will probably come from a different speaker or recording environment, compared to the training data. However, this differentiation is often not taken into account in the development of such models. So the question is: How important are the different voice characteristics of the speakers, when it comes to emotion recognition?

Sources of voice variability between speakers can be biological, such as gender and age, or socio-cultural, such as language, culture, and personal character. Indeed, due to the particular nature of emotion, its expression varies considerably, depending on the above. A first approach to reduce these differences is the normalization of data, using simple techniques. Of particular interest is a series of Speaker Adaptation techniques, developed in the field of Automatic Speech Recognition. In this work, we applied and compared the most basic ones, in order to recognize emotion. In addition, variants were investigated.

Another approach, given the different expression of the speakers, is finding their neutral speech. The knowledge of the neutral characteristics of each speaker makes it possible to detect any emotional charge, in order to improve the human-computer interaction. In this spirit, an existing system idea based on the bibliography has been developed. An important feature of the new system is the inclusion of a Speaker Adaptation technique.

Keywords: Speaker Adaptation, Emotion Recognition, Speech

Ευχαριστίες

Η παρούσα Διπλωματική Εργασία σηματοδοτεί την ολοκλήρωση των προπτυχιακών σπουδών μου στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών στο Εθνικό Μετσόβιο Πολυτεχνείο.

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον καθηγητή κ. Αλέξανδρο Ποταμιάνο για την εμπιστοσύνη του και την ευκαιρία που μου προσέφερε να ασχοληθώ με το συγκεκριμένο αντικείμενο μελέτης. Η υποστήριξη και οι συμβουλές του ήταν καθοριστικής σημασίας, σε όλα τα στάδια εκπόνησης της εργασίας. Επιπλέον, οι διαλέξεις του στα μαθήματα της *Επεξεργασίας Φωνής και Φυσικής Γλώσσας* και της *Αναγνώρισης Προτύπων* ήταν εκείνες που τράβηξαν το ενδιαφέρον μου προς αυτούς τους τομείς, και καθόρισαν σε μεγάλο βαθμό την πορεία των σπουδών μου.

Θα ήθελα να ευχαριστήσω, επίσης, την Αροδάμη Χωριανοπούλου για την παραχώρηση του κώδικα σχετικά με το *Affective Saliency Model*.

Ένα ιδιαίτερο ευχαριστώ, τέλος, θα ήθελα να εκφράσω προς τους γονείς μου και τον αδερφό μου για τη διαρκή υποστήριξη και την έμπνευση που μου παρέχουν, καθώς επίσης και προς τον Χρίστο Π. για την ουσιαστική παρουσία του σε ευχάριστες αλλά και σε δύσκολες στιγμές.

Στη γιαγιά μου, Ευαγγελία

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1

Εισαγωγή

1.1 Ανθρώπινη Επικοινωνία και Συναίσθημα

Η ανθρώπινη επικοινωνία αποτελεί ιδιαίτερα σημαντικό στοιχείο της κοινωνίας και των σχέσεων που αναπτύσσονται μεταξύ των ανθρώπων. Ένα αξιοσημείωτο χαρακτηριστικό της είναι ότι περιλαμβάνει τόσο λεκτική πληροφορία, όσο και μη λεκτική. Ο τόνος της φωνής, οι εκφράσεις του προσώπου, οι χειρονομίες και η στάση του σώματος είναι μόνο λίγα από τα στοιχεία εκείνα, τα οποία συνδυάζονται με τις λέξεις για να δώσουν κάποια πληροφορία στο συνομιλητή. Μάλιστα, είναι δυνατή η διάκριση δύο καναλιών στην ανθρώπινη αλληλεπίδραση, σύμφωνα με το [1]: το άμεσο (*explicit*) και το έμμεσο (*implicit*) κανάλι. Το πρώτο μεταδίδει σαφή μηνύματα, ενώ το δεύτερο πληροφορεί τους ανθρώπους πώς να αντιληφθούν τα μηνύματα αυτά.

Ανάμεσα στα μη λεκτικά μηνύματα της ανθρώπινης επικοινωνίας, που μεταδίδονται μέσω του έμμεσου καναλιού, ιδιαίτερο ενδιαφέρον παρουσιάζει η συναισθηματική κατάσταση του ομιλητή. Ένα μικρό παιδί μπορεί να αντιληφθεί συναισθήματα, πριν αρχίσει να καταλαβαίνει το περιεχόμενο της ομιλίας [2]. Το συναίσθημα που βιώνει ένας ομιλητής, καθορίζει σε μεγάλο βαθμό τόσο τον τρόπο ομιλίας του, όσο και την αντίληψη των μηνυμάτων από την άλλη πλευρά. Επηρεάζει, επίσης, όλα τα προαναφερθέντα στοιχεία της ανθρώπινης αλληλεπίδρασης, όπως τον τόνο και την ένταση της φωνής ή τις εκφράσεις του προσώπου, μέσω των οποίων επιδιώκει να δώσει την αντίστοιχη πληροφορία. Ο Beethoven, αφού είχε γίνει κουφός, έγραψε ότι μπορούσε να κρίνει από την έκφραση του προσώπου ενός εκτελεστή, αν ερμήνευε το μουσικό κομμάτι του στο σωστό πνεύμα [2]. Τέλος, ιδιαίτερο χαρακτηριστικό της φύσης του συναισθήματος αποτελεί η διαφορετική έκφρασή του μεταξύ των ανθρώπων, η οποία επηρεάζεται από το χαρακτήρα, αλλά και από το κοινωνικό-πολιτισμικό περιβάλλον του κάθε ατόμου.

1.2 Αναγνώριση Συναισθήματος στην Τεχνολογία

Εφόσον το συναίσθημα αποτελεί τόσο σημαντικό στοιχείο της ανθρώπινης επικοινωνίας, κρίνεται αναγκαία η έρευνά του, με σκοπό τη βελτίωση της αλληλεπίδρασης ανθρώπου-μηχανής. Ένα παράδειγμα που δείχνει τη σημασία του συναισθήματος σε τέτοια περίπτωση αλληλεπίδρασης είναι το *Turing test* [2]. Το τεστ αυτό εξετάζει αν ένας άνθρωπος μπορεί να αναγνωρίσει την πηγή των απαντήσεων ενός υπολογιστή, κατά την επικοινωνία με αυτόν, αν δηλαδή προέρχονται από έναν άνθρωπο ή από το ίδιο το μηχάνημα. Ο άνθρωπος μπορεί να συζητήσει για οποιο-

δήποτε θέμα με τον υπολογιστή, περιγράφοντας για παράδειγμα ένα ευχάριστο ή δυσάρεστο γεγονός. Τότε, οι απαντήσεις του υπολογιστή θα πρέπει να μην μπορούν να διαχωριστούν από τις αντίστοιχες πιθανές απαντήσεις ενός ανθρώπου συνομιλητή, έτσι ώστε να περάσει το τεστ. Προφανώς, δεν είναι δυνατόν ο υπολογιστής να περάσει το τεστ αυτό, αν δεν έχει την ικανότητα αντίληψης και έκφρασης συναισθημάτων.

Τα τελευταία χρόνια έχει σημειωθεί σημαντική έρευνα όσον αφορά την υπολογιστική αναγνώριση συναισθήματος (*Affective Computing*). Εξάγοντας χαρακτηριστικά από φωνή, βίντεο ή κείμενο, επιδιώκεται η αντιστοίχισή τους σε διακριτό ή συνεχή χώρο συναισθημάτων. Σκοπός είναι η ανάπτυξη της ικανότητας των υπολογιστών να ανιχνεύουν ένα συναίσθημα και να αντιδρούν κατάλληλα. Μάλιστα, στο [2] γίνεται αναφορά σε 4 περιπτώσεις ικανοτήτων, με βάση το συνδυασμό: αν ο υπολογιστής μπορεί να αντιληφθεί ή όχι συναισθήματα και αν μπορεί αντίστοιχα να εκφράσει ή όχι συναισθήματα. Όπως είναι λογικό, η ανάπτυξη τέτοιων ικανοτήτων σε μηχανές, απαιτεί συνεργασία και μελέτη πολλών κλάδων ταυτόχρονα, όπως Νευρολογία, Ψυχολογία και Γνωσιακή επιστήμη, εκτός από την πληροφορική. Σε αυτό το πνεύμα, στο [3] υποστηρίζεται ότι για να πετύχει μια τέτοια προσπάθεια, είναι απαραίτητη η ένταξη ιδιαιτεροτήτων των ανθρώπινων νοητικών διεργασιών στα αναπτυσσόμενα μοντέλα.

Οι εφαρμογές της υπολογιστικής αναγνώρισης συναισθήματος είναι ποικίλες. Βασικό παράδειγμα αποτελούν τα διαλογικά συστήματα σε τηλεφωνικά κέντρα, με σκοπό τη βελτίωση της εξυπηρέτησης των πελατών και την αποφυγή αρνητικών συναισθημάτων, όπως θυμός ή εκνευρισμός. Επίσης, ιδιαίτερα χρήσιμη κρίνεται η αναγνώριση συναισθήματος σε συστήματα αυτόματης διδασκαλίας, τα οποία θα πρέπει να αντιλαμβάνονται αν τα παραδείγματα είναι βαρετά ή δύσκολα για το χρήστη [1]. Κομβικής σημασίας θα είναι η συνεισφορά της αναγνώρισης συναισθήματος στην ψυχιατρική κατά τη διάγνωση ασθενειών, όπως η κατάθλιψη, η ασθένεια του Parkinson και η διπολική διαταραχή [4, 5]. Άλλες εφαρμογές αφορούν την ψυχαγωγία, την ανάλυση κειμένου και την προώθηση προϊόντων (marketing). Τέλος, με την ανάπτυξη των ρομπότ τα τελευταία χρόνια, είναι απαραίτητη η αντίληψη συναισθήματος από αυτά, έτσι ώστε να είναι εφικτή οποιαδήποτε αλληλεπίδραση με τον άνθρωπο.

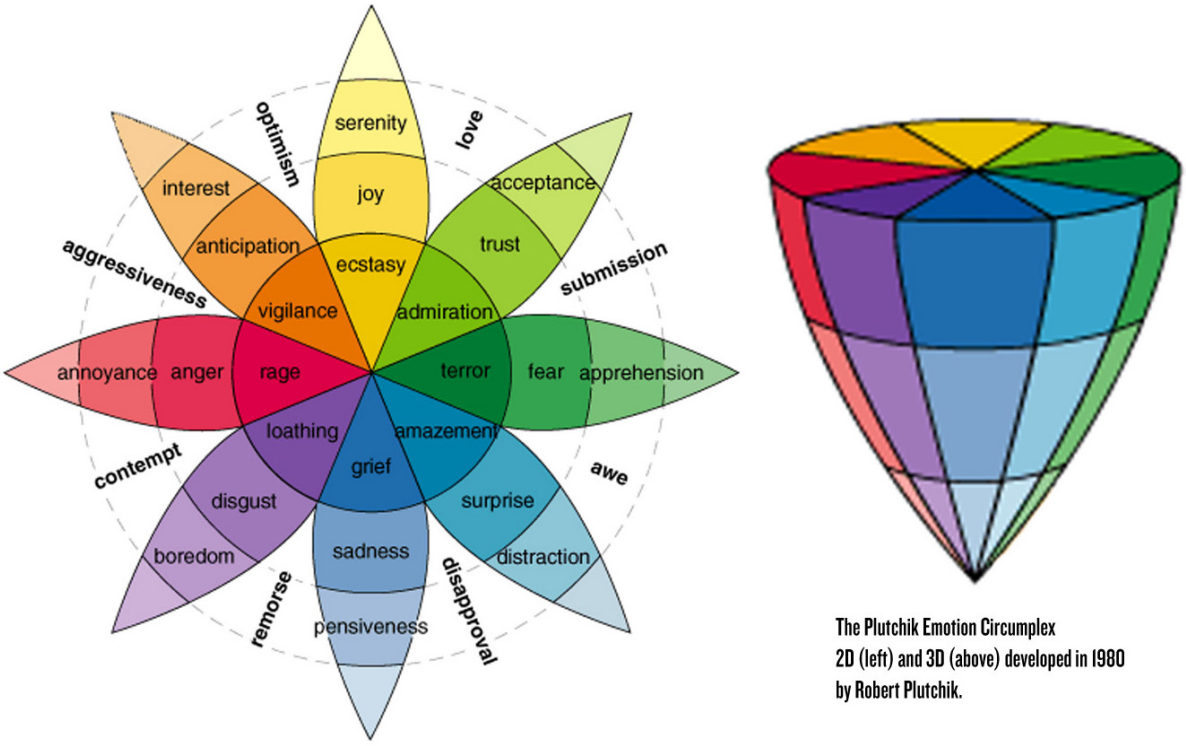
Στην πράξη, η αναγνώριση συναισθήματος αντιμετωπίζεται ως πρόβλημα Αναγνώρισης Προτύπων. Αρχικά, προσδιορίζονται οι κλάσεις στις οποίες θα πρέπει να ανήκουν τα δεδομένα. Οι κλάσεις είναι συναισθηματικές και αντιστοιχούν είτε σε διακριτές ετικέτες είτε σε συνεχές τιμές, όπως αναλύεται παρακάτω. Ακολουθεί η εξαγωγή των χαρακτηριστικών, τα οποία περιγράφουν κάθε δεδομένο με χρήση διανυσμάτων. Το παρόν σύγγραμμα αναφέρεται σε Αναγνώριση Συναισθήματος από Φωνή, και άρα σε ακουστικά χαρακτηριστικά. Στη συνέχεια, είναι απαραίτητη η ανάπτυξη μοντέλου, όπου επιδιώκεται η όσο το δυνατόν καλύτερη μοντελοποίηση και ταξινόμηση των δεδομένων, με βάση τα χαρακτηριστικά τους. Έχει γίνει μεγάλη προσπάθεια και έρευνα σχετικά με τη καταλληλότητα των μοντέλων στην Αναγνώριση Συναισθήματος από Φωνή, ξεκινώντας από μοντέλα που χρησιμοποιούνται ήδη στην Αναγνώριση Φωνής ή και σε άλλους τομείς.

1.3 Κατηγοριοποίηση και Αναπαράσταση Συναισθημάτων

Πλήθος μοντέλων έχουν αναπτυχθεί με σκοπό την κατηγοριοποίηση και την αναπαράσταση των συναισθημάτων στην Ψυχολογία. Βασική προσέγγιση αποτελεί η χρήση διακριτών κατηγοριών.

Σε αυτό το πνεύμα, έχουν προταθεί μια σειρά από λίστες σχετικά με το ποια είναι τα βασικά συναισθήματα. Μια ευρέως διαδεδομένη κατηγοριοποίηση περιλαμβάνει τα εξής 6: φόβος (fear), θυμός (anger), χαρά (happiness), λύπη (sadness), έκπληξη (surprise) και αηδία (disgust) [6].

Θεωρώντας τα συναισθήματα τα οποία δεν είναι βασικά ως δευτερεύοντα, προκύπτουν διάφορες αναπαραστάσεις. Ευρέως γνωστή είναι αυτή του Plutchik [7], η οποία περιγράφει τις σχέσεις μεταξύ των συναισθημάτων μέσω της έννοιας του χρώματος και των αναμίξεών του, σύμφωνα με έναν τροχό χρωμάτων (βλ. Σχήμα 1.1). Τα 8 βασικά συναισθήματα, σύμφωνα με τη θεωρία αυτή, απεικονίζονται στον κύκλο, στο κέντρο του δισδιάστατου μοντέλου. Τα συναισθήματα στις άσπρες περιοχές προκύπτουν από τη μίξη δύο βασικών συναισθημάτων. Στο τρισδιάστατο μοντέλο, η κατακόρυφη διάσταση του κώνου αναπαριστά την ένταση και ο κύκλος αναπαριστά το βαθμό ομοιότητας των συναισθημάτων. Διακρίνονται 8 τομείς, οι οποίοι αντιστοιχούν στις βασικές συναισθηματικές διαστάσεις, ως τέσσερα ζεύγη αντιθέτων.

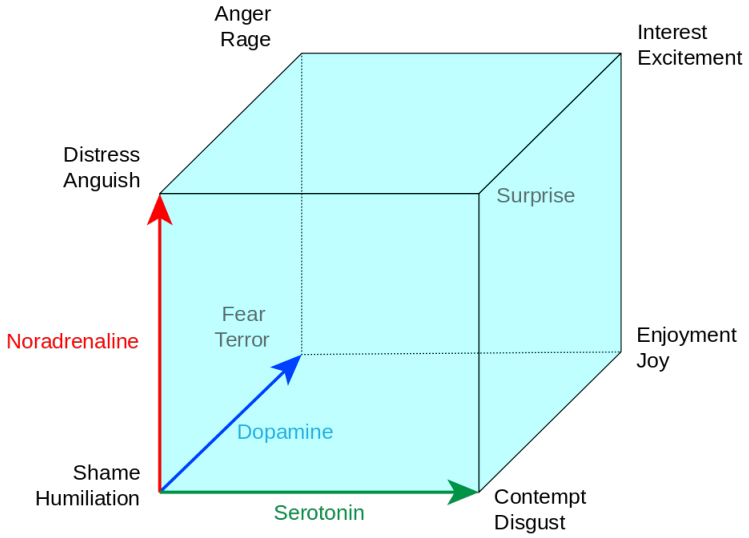


Σχήμα 1.1: *Αναπαράσταση συναισθημάτων του Plutchik. (Η εικόνα προέρχεται από το [7].)*

Εκτός από τη διακριτή κατηγοριοποίηση των συναισθημάτων, μια άλλη προσέγγιση είναι η αναπαράστασή τους πάνω σε δισδιάστατο ή τρισδιάστατο σύστημα αξόνων, με συνεχείς τιμές. Ιδιαίτερα γνωστό μοντέλο είναι το λεγόμενο *Pleasure-Arousal-Dominance (PAD) Emotional State Model* των Mehrabian και Russell [8], το οποίο περιλαμβάνει 3 άξονες: *Pleasure*, *Arousal* και *Dominance*. Ο πρώτος άξονας μετράει πόσο ευχάριστο είναι ένα συναίσθημα, ο δεύτερος

μετράει την ένταση του συναισθήματος και ο τρίτος αντιστοιχεί στη φύση του συναισθήματος αν είναι κυρίαρχη. Για παράδειγμα, ο θυμός είναι κυρίαρχο συναίσθημα σε σχέση με το φόβο, αλλά και τα δύο είναι δυσάρεστα. Επίσης, η πλήξη έχει χαμηλότερη τιμή έντασης από το θυμό. Συχνά, οι παραπάνω άξονες συναντούνται ως *Valence*, *Activation* και *Dominance* αντίστοιχα.

Ένα νεότερο μοντέλο τριών αξόνων, με σκοπό την αναπαράσταση 8 βασικών συναισθημάτων, είναι ο κύβος του Lövheim [9] (βλ. Σχήμα 1.2). Βασίζεται στους 3 νευροδιαβιβαστές μονοαμίνης: *Serotonin*, *Noradrenaline* και *Dopamine*, οι οποίοι παράγονται από λίγους νευρώνες στον εγκέφαλο και θεωρούνται ότι επηρεάζουν σε σημαντικό βαθμό τα συναισθήματα και τη συμπεριφορά του ανθρώπου. Σύμφωνα με το μοντέλο, για παράδειγμα, ο θυμός παράγεται από συνδυασμό χαμηλής Serotonin, υψηλής Dopamine και υψηλής Noradrenaline. Επίσης, αναφέρεται ότι συγκριτικά με παλαιότερα μοντέλα, ο συγκεκριμένος κύβος είναι ελαφρώς περιστραμμένος, αφού η διάσταση Valence, δεν αντιστοιχεί ούτε στη Serotonin, ούτε στη Dopamine.



Σχήμα 1.2: *Ο κύβος συναισθημάτων του Lövheim (2011). (Η εικόνα προέρχεται από το [9].)*

1.4 Διαφορετικά Στοιχεία των Ομιλητών

Γενικά, κάθε ομιλητής μιλάει με το δικό του τρόπο. Στοιχεία που διαφοροποιούν τη φωνή του από τους υπόλοιπους ομιλητές μπορεί να είναι βιολογικά, όπως το φύλο και η ηλικία, ή κοινωνικο-πολιτισμικά, όπως η γλώσσα. Επίσης, όπως αναφέρθηκε και παραπάνω, σημαντική επιρροή στο σήμα φωνής έχει και η έκφραση συναισθήματος. Όμως, η έκφραση αυτή δεν είναι η ίδια μεταξύ των ομιλητών, αφού κάθε ένας από αυτούς χαρακτηρίζεται από τον προσωπικό του χαρακτήρα, την κουλτούρα του και άλλα ιδιαίτερα στοιχεία. Όλα τα παραπάνω συμβάλλουν στη διαφοροποίηση του σήματος φωνής κάθε ατόμου, και άρα στα ακουστικά χαρακτηριστικά που εξάγονται από κάθε εκφώνηση. Τα χαρακτηριστικά αυτά θα χρησιμοποιηθούν, στη συνέχεια, για την εκπαίδευση ή αξιολόγηση ενός συστήματος αναγνώρισης συναισθήματος από φωνή. Έτσι, η οποιαδήποτε διαφοροποίηση μεταξύ των ομιλητών θα οδηγήσει πιθανότητα σε χαμηλή

απόδοση του συστήματος, αφού αυξάνεται η πιθανότητα λανθασμένης αναγνώρισης.

Για παράδειγμα, δύο άνθρωποι που βιώνουν το ίδιο συναίσθημα, όπως η λύπη, θα το εκφράσουν πιθανότατα με διαφορετικό τρόπο, ο οποίος εξαρτάται από το βαθμό της λύπης τους, το χαρακτήρα τους, αλλά και το κοινωνικο-πολιτισμικό περιβάλλον, μέσα στο οποίο έχουν μάθει να εκφράζονται. Επίσης, ασάφεια μπορεί να δημιουργηθεί και από την ομοιότητα των ακουστικών χαρακτηριστικών δύο διαφορετικών συναισθημάτων, όπως λύπη και πλήξη. Επιπλέον, ως στοιχείο διαφοροποίησης μπορεί να θεωρηθεί και το περιβάλλον ηχογράφησης κάθε ομιλητή, καθώς επιφέρει μεταβολές στο τελικό σήμα φωνής. Έτσι, σημαντικό ρόλο μπορεί να παίξει η αλλαγή της στάθμης της ενέργειας του σήματος φωνής εξαιτίας διαφορετικών συνθηκών ηχογράφησης, καθώς η αύξηση της ενέργειας είναι αυτή που χαρακτηρίζει κάποια είδη συναισθήματος, όπως θυμό ή ενθουσιασμό.

Όλα τα παραπάνω παραδείγματα δείχνουν μεταβολές που μπορεί να παρατηρηθούν σε ένα σήμα φωνής και αντίστοιχα να επηρεάσουν τα εξαγόμενα ακουστικά χαρακτηριστικά. Τέτοιες μεταβολές είναι γενικά ανεκτές από το ανθρώπινο σύστημα αναγνώρισης, καθώς εκπαιδεύεται διαρκώς σε ποικιλία ομιλητών στις καθημερινές αλληλεπιδράσεις. Επίσης, λόγω της πολυπλοκότητάς του, ο ανθρώπινος εγκέφαλος μπορεί να αντιλαμβάνεται πλήθος καινούργιων ερεθισμάτων. Όμως, δεν ισχύει το ίδιο για ένα αυτόματο σύστημα αναγνώρισης, ανεπτυγμένο σε έναν υπολογιστή. Αυτό, εκπαιδεύεται με βάση κάποιον περιορισμένο αριθμό δεδομένων και στη συνέχεια, καλείται να αναγνωρίσει το συναίσθημα ενός δεδομένου εισόδου, το οποίο συχνά θα προέρχεται από διαφορετικό ομιλητή ή περιβάλλον ηχογράφησης, συγκριτικά με τα δεδομένα εκπαίδευσης. Με αυτόν τον τρόπο, γίνεται κατανοητό ότι η διαφοροποίηση της φωνής μεταξύ των ομιλητών αποτελεί κρίσιμο στοιχείο της απόδοσης ενός συστήματος, ειδικά σε πραγματικές εφαρμογές όπου τα ηχητικά δεδομένα προέρχονται από διαφορετικούς ομιλητές ή ακουστικά περιβάλλοντα. Εκτενέστερη ανάλυση των διαφορετικών στοιχείων των ομιλητών παρατίθεται στην παράγραφο 2.5.

1.5 Συνεισφορά της Εργασίας

Η συγκεκριμένη εργασία επικεντρώνεται στην προσαρμογή ενός ομιλητή σε ένα σύστημα αναγνώρισης συναισθήματος από φωνή, με σκοπό τη μείωση των διαφορών που αναλύθηκαν παραπάνω. Η μείωση των διαφορών αυτών θα οδηγήσει σε ανεκτικότητα του συστήματος σε διαφορετικούς ομιλητές ή συνθήκες ηχογράφησης, με αποτέλεσμα την αύξηση της απόδοσής του. Μια πρώτη προσέγγιση για τη μείωση των διαφορών αυτών είναι η κανονικοποίηση των δεδομένων, με χρήση απλών τεχνικών. Ιδιαίτερο ενδιαφέρον παρουσιάζουν μια σειρά από τεχνικές Προσαρμογής του Ομιλητή, που έχουν αναπτυχθεί στον τομέα της Αυτόματης Αναγνώρισης Φωνής, με σκοπό τη βελτίωση της επίδοσης ενός μοντέλου. Μικρή προσπάθεια έχει γίνει, μέχρι τώρα, για την εφαρμογή των τεχνικών αυτών στον τομέα της Αναγνώρισης Συναισθήματος από Φωνή. Εδώ, γίνεται εφαρμογή και αξιολόγηση της αποτελεσματικότητας των πιο βασικών από αυτές. Επιπλέον, επιχειρούνται κάποιες παραλλαγές τους με επιτυχία.

Μια άλλη προσέγγιση, δοθείσας της διαφορετικής έκφρασης των ομιλητών, βασίζεται στην εύρεση της ουδέτερης ομιλίας τους. Στόχος είναι, δηλαδή, ο διαχωρισμός των συναισθηματικών φράσεων από τις ουδέτερες, προσδιορίζοντας αρχικά τα ουδέτερα χαρακτηριστικά του συγκεκριμένου ομιλητή. Αυτή η προσέγγιση μπορεί να συνεισφέρει στη βελτίωση των δυνατοτήτων μιας πραγματικής εφαρμογής που περιλαμβάνει αλληλεπίδραση ανθρώπου-μηχανής, καθώς ανι-

χνεύει τη συναισθηματική φόρτιση ενός ατόμου. Σε αυτό το πνεύμα, στη συγκεκριμένη εργασία εξελίσσεται μια ήδη ανεπτυγμένη ιδέα συστήματος με βάση τη βιβλιογραφία, εντάσσοντας και τεχνική Προσαρμογής του Ομιλητή. Τα αποτελέσματα δείχνουν σημαντική βελτίωση.

1.6 Περίγραμμα της Εργασίας

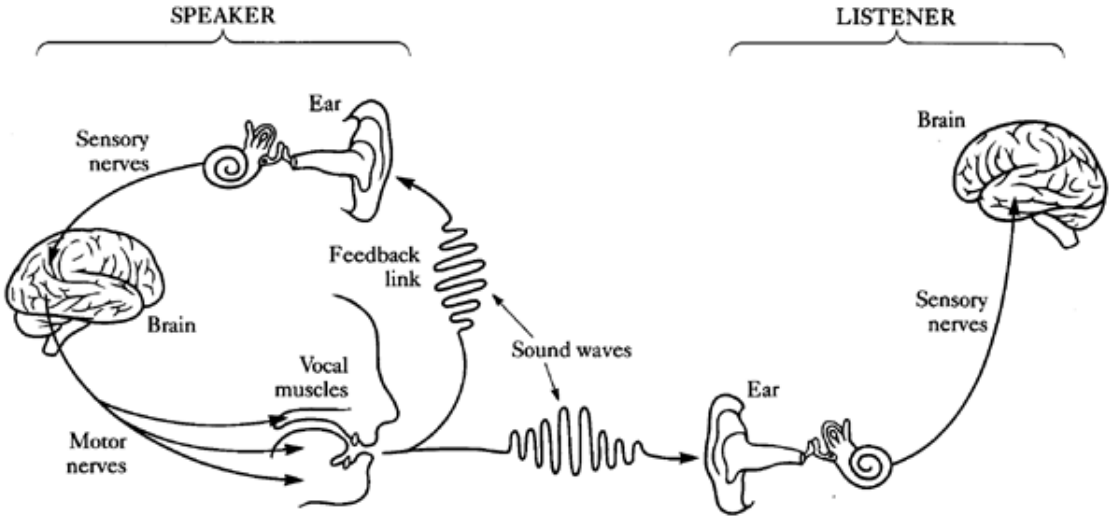
Η οργάνωση της συγκεκριμένης εργασίας συνοψίζεται ως ακολούθως. Στο Κεφάλαιο 2 παρουσιάζεται ο μηχανισμός παραγωγής φωνής του ανθρώπου, καθώς και τα εξαγόμενα ακουστικά χαρακτηριστικά και η σχέση τους με το συναίσθημα. Επίσης, περιγράφονται τα διαφορετικά στοιχεία μεταξύ των ομιλητών. Στο Κεφάλαιο 3 αναλύονται κάποιοι βασικοί ταξινομητές και μοντέλα, που αξιοποιούνται σε επόμενα κεφάλαια. Στο Κεφάλαιο 4 παρουσιάζονται κάποιες τεχνικές κανονικοποίησης των χαρακτηριστικών, καθώς γίνεται και εφαρμογή τους. Στο Κεφάλαιο 5 αναλύονται οι βασικές τεχνικές Προσαρμογής του Ομιλητή, που έχουν αναπτυχθεί στον τομέα της Αυτόματης Αναγνώρισης Φωνής. Στο Κεφάλαιο 6 παρουσιάζεται η εφαρμογή τους, με σκοπό τη βελτίωση της Αναγνώρισης Συναισθήματος από Φωνή, καθώς επίσης προτείνεται μια παραλλαγή με χρήση διαφορετικού μοντέλου. Στο Κεφάλαιο 7 περιγράφεται μια προσέγγιση ανίχνευσης της ουδέτερης ομιλίας ενός ομιλητή, με ένταξη τεχνικής προσαρμογής. Τέλος, η εργασία ολοκληρώνεται με τα συμπεράσματα και τις μελλοντικές προεκτάσεις του Κεφαλαίου 8.

Κεφάλαιο 2

Χαρακτηριστικά Φωνής

2.1 Εισαγωγή

Η διαδικασία παραγωγής και αντίληψης της ομιλίας, κατά την επικοινωνία δύο ανθρώπων, μπορεί να απεικονιστεί σχηματικά με την αλυσίδα ομιλίας (*speech chain*) του Σχήματος 2.1. Η διαδικασία ξεκινά με την αναπαράσταση ενός μηνύματος στον εγκέφαλο του ομιλητή. Μετά τη διαμόρφωση του μηνύματος αυτού σε κείμενο, μετατρέπεται μέσω του κώδικα της γλώσσας σε μία ακολουθία φωνητικών συμβόλων, όπου περιλαμβάνεται και η πληροφορία της συναισθηματικής φόρτισης [10]. Ακολουθεί η μετατροπή των φωνητικών συμβόλων σε σήματα *νευρομυικού ελέγχου*, τα οποία ορίζουν την κίνηση των αρθρωτών ομιλίας, δηλαδή της γλώσσας, του σαγονιού, των χειλιών, των δοντιών και του ουρανίσκου. Η κίνηση αυτή οδηγεί στην παραγωγή των επιθυμητών ήχων κατά την ομιλία, καθώς η ροή αέρα εκφέρεται μέσω της φωνητικής οδού του



Σχήμα 2.1: *Η αλυσίδα ομιλίας. (Η εικόνα προέρχεται από το [11].)*

ομιλητή. Τα ηχητικά κύματα που δημιουργούνται φτάνουν στο αυτί του ακροατή. Εκεί, γίνεται η φασματική τους αναπαράσταση, αναλύοντας τις φασματικές συνιστώσες του εισερχόμενου σήματος με έναν μηχανισμό, που μοιάζει με μη ομοιόμορφη συστοιχία φίλτρων. Στη συνέχεια,

τα φασματικά χαρακτηριστικά οδηγούνται στον εγκέφαλο του ακροατή, όπου αποκωδικοποιούνται σε ένα σύνολο φωνημάτων, λέξεων και προτάσεων. Με αυτόν τον τρόπο, ο ακροατής αντιλαμβάνεται το νόημα του μηνύματος φωνής. Παράλληλα, στο σχήμα παρατηρείται μία ζεύξη ανάδρασης, όπου τα ηχητικά κύματα φτάνουν με τον ίδιο τρόπο και στο αυτί του ομιλητή.

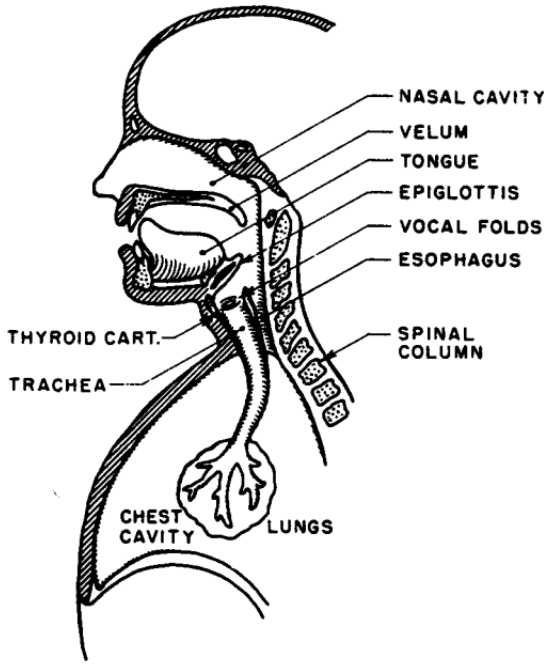
Η παραπάνω διαδικασία περιγράφει συνοπτικά τα στάδια της επικοινωνίας των ανθρώπων, κατά την ομιλία τους. Βασικό στοιχείο, το οποίο ενδιαφέρει στην παρούσα εργασία, είναι η φωνή. Η φωνή του ανθρώπου παρουσιάζει ιδιαίτερο ενδιαφέρον, τόσο ως προς το μηχανισμό παραγωγής της, όσο και ως προς τα χαρακτηριστικά της. Αποτελεί ένα ηχητικό σήμα, το οποίο έχει ως σκοπό τη μετάδοση κάποιου είδους πληροφορίας. Όπως αναφέρθηκε και στο Κεφάλαιο 1, η πληροφορία αυτή είναι η σύνθεση των σαφών λεκτικών μηνυμάτων με τα έμμεσα στοιχεία, μεταξύ των οποίων και το συναίσθημα. Το συναίσθημα μεταβάλλει τα χαρακτηριστικά της φωνής, όπως τον τόνο ή την ένταση, επηρεάζοντας με αυτόν τον τρόπο τη σημασία του συνολικού μηνύματος, και άρα την επικοινωνία των ανθρώπων.

Παρακάτω αναλύεται ο μηχανισμός παραγωγής της φωνής του ανθρώπου, καθώς και τα ακουστικά χαρακτηριστικά που εξάγονται από το σήμα φωνής με σκοπό την περιγραφή κάθε δεδομένου. Επίσης, παρουσιάζεται η επιρροή της συναισθηματικής έκφρασης στα χαρακτηριστικά αυτά. Τέλος, γίνεται αναφορά στα διαφοροποίηση τους ανάλογα με τον ομιλητή. Τα διαφορετικά στοιχεία που παρατηρούνται μεταξύ των ομιλητών, μεταβάλλουν τα εξαγόμενα ακουστικά χαρακτηριστικά, με αποτέλεσμα να επηρεάζουν την απόδοση ενός συστήματος αναγνώρισης συναισθήματος.

2.2 Παραγωγή Φωνής στον Άνθρωπο

Στο Σχήμα 2.2 απεικονίζεται σχηματικά το ανθρώπινο σύστημα φωνής, περιλαμβάνοντας τα μέρη του σώματος που συμμετέχουν κατά την παραγωγή ομιλίας. Αρχικά, οι πνεύμονες και η θωρακική κοιλότητα συμβάλλουν στην ώθηση αέρα προς τη φωνητική οδό, με σκοπό τη διέγερσή της. Ο αέρας αυτός, αφού έχει εισέλθει στους πνεύμονες με τη φυσιολογική αναπνοή, κατευθύνεται προς τις φωνητικές χορδές, μέσω της τραχείας. Λόγω της ροής αέρα, οι φωνητικές χορδές πάλλονται, ανοίγοντας και κλείνοντας περιοδικά. Ο αέρας, στη συνέχεια, φτάνει στη φωνητική οδό, η οποία ξεκινά από το άνοιγμα ανάμεσα στις φωνητικές χορδές (τη γλωττίδα) και τελειώνει στα χείλη. Ουσιαστικά, η φωνητική οδός αποτελείται από το φάρυγγα (τμήμα που συνδέει τον οισοφάγο με το στόμα), τη στοματική κοιλότητα (γλώσσα, χείλη, σαγόνι, στόμα), και κατά περίπτωση τη ρινική κοιλότητα (ανάλογα με τη θέση της μεμβράνης του ουρανίσκου) [10]. Η ρινική κοιλότητα ξεκινά από τη μεμβράνη του ουρανίσκου και καταλήγει στα ρουθούνια. Για την παραγωγή έρρινων ήχων, η μεμβράνη αυτή κατεβαίνει, με σκοπό τη σύνδεση της ρινικής οδού με τη φωνητική.

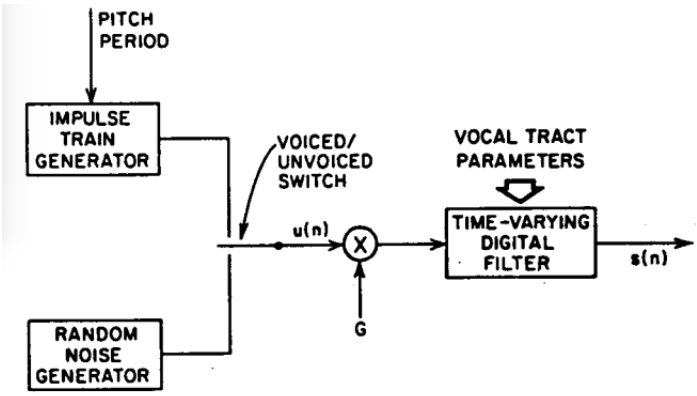
Όσον αφορά την παραγωγή *έμφωνων* ήχων, όπως τα φωνήεντα, παρατηρείται ροή του αέρα, η οποία διακόπτεται περιοδικά από το άνοιγμα και το κλείσιμο της γλωττίδας [10]. Οι φωνητικές χορδές πάλλονται σύμφωνα με μία ταλάντωση χαλάρωσης, παράγοντας έτσι σχεδόν περιοδικούς παλμούς αέρα. Η θέση, το σχήμα και το μέγεθος των διάφορων αρθρωτών (χείλη, δόντια, γλώσσα, σαγόνι, μεμβράνη ουρανίσκου) μεταβάλλονται, με σκοπό την παραγωγή των επιθυμητών ήχων. Οι *άφωνοι* ήχοι, όπως τα /σ/, /τ/, /π/, παράγονται κατά τη δημιουργία μερικής σύσφιξης σε κάποιο σημείο της φωνητικής οδού, συνήθως προς το άκρο της στοματικής κοιλότητας. Τότε, η ώθηση του αέρα γίνεται με μεγάλη ταχύτητα, αρκετή ώστε να παραχθεί ο



Σχήμα 2.2: *Σχηματική αναπαράσταση της ανθρώπινης φωνητικής οδού. (Η εικόνα προέρχεται από το [12].)*

επιθυμητός ήχος. Ουσιαστικά, η διέγερση της φωνητικής οδού στους άφωνους ήχους, γίνεται μέσω πηγής θορύβου.

Είναι δυνατή η μοντελοποίηση του μηχανισμού παραγωγής φωνής σύμφωνα με το Σχήμα 2.3. Στο σχήμα αυτό, ένας διακόπτης ορίζει αν ο παραγόμενος ήχος είναι έμφωνος ή άφωνος. Στην πρώτη περίπτωση η διέγερση είναι μια ακολουθία κρουστικών παλμών, με συγκεκριμένη περίοδο (θεμελιώδης περίοδος). Το αντίστροφο της θεμελιώδους περιόδου ονομάζεται θεμελιώδης συχνότητα ή *pitch*, και χρησιμοποιείται συχνά ως ακουστικό χαρακτηριστικό. Στη δεύτερη περίπτωση, σε αυτή δηλαδή των άφωνων ήχων, η διέγερση είναι μία πηγή τυχαίου θορύβου. Το σήμα, που παράγεται, ενισχύεται με κέρδος G, και στη συνέχεια συνελίσσεται με τη κρουστική απόκριση της φωνητικής οδού, όπου εδώ συμβολίζεται με ένα χρονικά μεταβαλλόμενο ψηφιακό



Σχήμα 2.3: *Μοντέλο παραγωγής φωνής. (Η εικόνα προέρχεται από το [12].)*

φίλτρο. Το συγκεκριμένο σχήμα χρησιμοποιείται και για τη σύνθεση ομιλίας, με βάση τη μέθοδο Γραμμικής Πρόβλεψης (Linear Prediction Coding ή LPC). Η ανάλυση Γραμμικής Πρόβλεψης αποτελεί μία ευρέως διαδεδομένη τεχνική ανάλυσης φωνής, καθώς μπορεί να εκτιμήσει τις κύριες παραμέτρους παραγωγής φωνής. Η βασική ιδέα είναι η προσέγγιση ενός δείγματος φωνής με γραμμικό συνδυασμό p παρελθόντων δειγμάτων [10]. Με ελαχιστοποίηση του αθροίσματος των τετραγώνων των διαφορών μεταξύ των πραγματικών δειγμάτων φωνής και των προβλέψεων, προσδιορίζεται ένα σύνολο p συντελεστών πρόβλεψης. Οι συντελεστές αυτοί αφορούν τους συντελεστές του ψηφιακού φίλτρου όλο-πόλων, το οποίο αναπαριστά το σύστημα της φωνητικής οδού.

2.3 Ακουστικά Χαρακτηριστικά

Πριν την οποιαδήποτε χρήση των ηχητικών δεδομένων, με σκοπό την ανάπτυξη ή αξιολόγηση ενός συστήματος, είναι απαραίτητη η εξαγωγή ακουστικών χαρακτηριστικών από αυτά. Τα χαρακτηριστικά αυτά είναι οι λεγόμενοι περιγραφητές, που κάνουν εφικτή τη μοντελοποίηση κάθε δεδομένου. Στην Αναγνώριση Συναισθήματος από Φωνή έχει χρησιμοποιηθεί μια ποικιλία ακουστικών χαρακτηριστικών, πολλά από τα οποία προέρχονται από την Αναγνώριση Φωνής. Σκοπός είναι η ανίχνευση όλων εκείνων των στοιχείων, που επηρεάζονται από τη συναισθηματική έκφραση. Τα στοιχεία αυτά και οι μεταβολές τους θα οδηγήσουν στην όσο το δυνατόν ορθότερη αναγνώριση κάθε συναισθήματος.

Θεωρώντας ότι κάθε δεδομένο αντιστοιχεί σε μία ηχητική εκφώνηση ενός ομιλητή, η εξαγωγή χαρακτηριστικών μπορεί να γίνει σε επίπεδο πλαισίου, ομάδας πλαισίων ή και εκφώνησης. Σύμφωνα με το [13], τα διανύσματα χαρακτηριστικών διακρίνονται σε 2 κατηγορίες: *short-time* και *long-time*. Η πρώτη κατηγορία αντιστοιχεί στα εξαγόμενα χαρακτηριστικά ανά πλαίσιο, όπου η εκφώνηση έχει χωριστεί σε πλαίσια ίσης διάρκειας (συνήθως 25-50 msec) με χρήση τεχνικών παραθύρωσης. Στη δεύτερη κατηγορία, αντίθετα, ανήκουν τα χαρακτηριστικά εκείνα που εξάγονται από σήμα μεγαλύτερης διάρκειας, ακόμα και από ολόκληρη την εκφώνηση.

Μια άλλη πιθανή κατηγοριοποίηση είναι ο διαχωρισμός χαμηλού-επιπέδου περιγραφητών (*low-level descriptors* ή *LLDs*) και συναρτησιακών (*functionals*). Οι πρώτοι περιγραφητές περιλαμβάνουν χαρακτηριστικά φασματικά, προσωδιακά και ποιότητας φωνής, καθώς και τις παραγώγους τους. Τα φασματικά αντιστοιχούν σε short-term χαρακτηριστικά, ενώ τα άλλα δύο σε long-term, σύμφωνα με την παραπάνω κατηγοριοποίηση [13]. Με βάση αυτούς τους χαμηλού-επιπέδου περιγραφητές, είναι δυνατός ο υπολογισμός στατιστικών, με σκοπό την παραγωγή χαρακτηριστικών ανά ομάδα πλαισίων ή ανά εκφώνηση. Τα στατιστικά αυτά αντιστοιχούν στα λεγόμενα συναρτησιακά (functionals).

Παρακάτω, δίνεται μια σύντομη περιγραφή των ακουστικών χαρακτηριστικών που χρησιμοποιούνται ευρέως στην Αναγνώριση Συναισθήματος από Φωνή. Ονομαστικά, αναγράφονται στον Πίνακα 2.1 συνήθεις χαμηλού-επιπέδου περιγραφητές.

	LLDs
Φασματικά	Mel-Frequency Cepstral Coefficients (MFCCs), Ενέργεια βραχέος χρόνου, Φασματική Ροή Linear Prediction Cepstral Coefficients (LPCCs), Συχνότητα Φωνοσυντονισμού (Formant), Πλάτος, Εύρος ζώνης συχνότητας φωνοσυντονισμού
Προσωδιακά	Θεμελιώδης Συχνότητα (Pitch), Ενέργεια, Ρυθμός Διέλευσης από το Μηδέν
Ποιότητα Φωνής	Harmonics-to-Noise Ratio (HNR), Shimmer, Jitter

Πίνακας 2.1: *Ακουστικά χαρακτηριστικά χαμηλού-επιπέδου (low-level descriptors).*

Θεμελιώδης Συχνότητα (Pitch)

Βασικό χαρακτηριστικό του σήματος φωνής είναι η θεμελιώδης συχνότητά του (*pitch*). Ουσιαστικά, αντιστοιχεί στη συχνότητα με την οποία ανοίγουν και κλείνουν οι φωνητικές χορδές κατά την παραγωγή του [14]. Συχνά, συμβολίζεται με F0, καθώς είναι η χαμηλότερη συχνότητα του σήματος και ακολουθούν οι συχνότητες φωνοσυντονισμού (*formants*) F1, F2 κλπ. Για τον υπολογισμό της θεμελιώδους συχνότητας, μια ευρέως διαδεδομένη μέθοδος βασίζεται στην αυτοσυσχέτιση του σήματος [14].

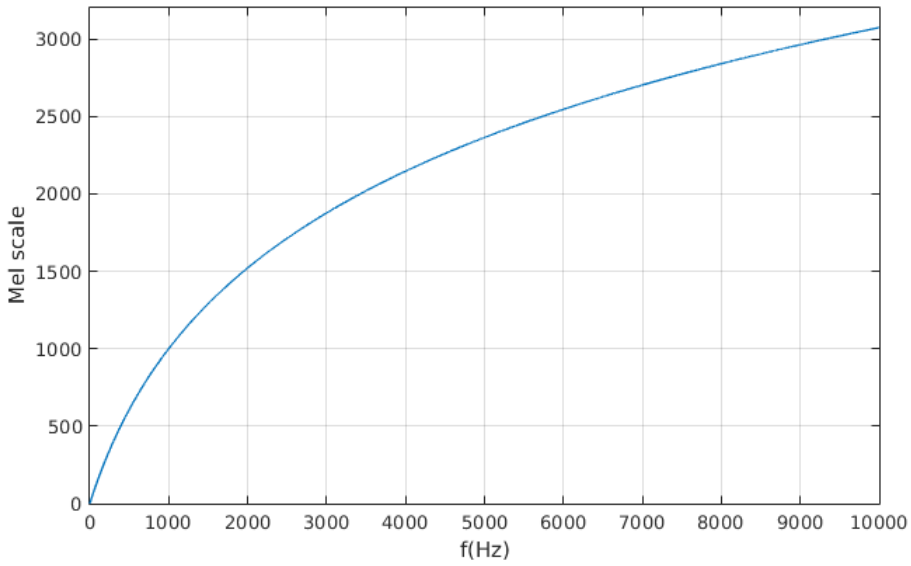
Στο αυτί του ανθρώπου, το pitch αντιστοιχεί με κάποιο τρόπο στο τονικό ύψος του σήματος [10]. Το αντιληπτό τονικό ύψος αποτελεί υποκειμενική ιδιότητα του ήχου, με μονάδα μέτρησης τα mel, ενώ η συχνότητα αποτελεί αντικειμενική ιδιότητα, με μονάδα μέτρησης τα Hz. Η κλίμακα mel (από τη λέξη melody, μελωδία) δίνει την υποκειμενική αντίληψη των συχνοτήτων από τον άνθρωπο, έτσι ώστε να είναι ισοκατανεμημένες σύμφωνα με την ακοή του. Επειδή, όμως, η διακριτική ικανότητα του ανθρώπου είναι μεγαλύτερη στις χαμηλές συχνότητες από τις υψηλές, η αντιστοιχία της κλίμακας mel με τις συχνότητες σε Hz προκύπτει λογαριθμική. Σύμφωνα με το [10], η συχνότητα f σε Hz υπολογίζεται στην κλίμακα mel με τη σχέση:

$$m = 1127 log_e(1 + \frac{f}{700}) \tag{2.3.1}$$

Η αντιστοιχία των δύο κλιμάκων φαίνεται στη γραφική παράσταση του Σχήματος 2.4.

Συχνότητα Φωνοσυντονισμού (Formant)

Στο χώρο των συχνοτήτων, οι συχνότητες φωνοσυντονισμού (formants) αντιστοιχούν στους συντονισμούς της φωνητικής οδού και σχετίζονται με τη μορφή και τις διαστάσεις της [14]. Κάθε μορφή της φωνητικής οδού χαρακτηρίζεται από ένα σύνολο τέτοιων συχνοτήτων. Η μορφή αυτή μεταβάλλεται με το χρόνο, με σκοπό τη διαμόρφωση του επιθυμητού ήχου κάθε φορά. Κάθε συχνότητα φωνοσυντονισμού χαρακτηρίζεται από μία κεντρική συχνότητα και ένα εύρος ζώνης. Εφόσον το σχήμα της φωνητικής οδού επηρεάζεται και από τη συναισθηματική κατάσταση του ομιλητή, οι συγκεκριμένες συχνότητες είναι ιδιαίτερα χρήσιμα χαρακτηριστικά για την αναγνώριση συναισθήματος. Για παράδειγμα, έχει βρεθεί ότι τα άτομα σε κατάσταση άγχους ή κατάθλιψης δεν αρθρώνουν έμφωνους ήχους με τον ίδιο τρόπο, όπως στην ουδέτερη

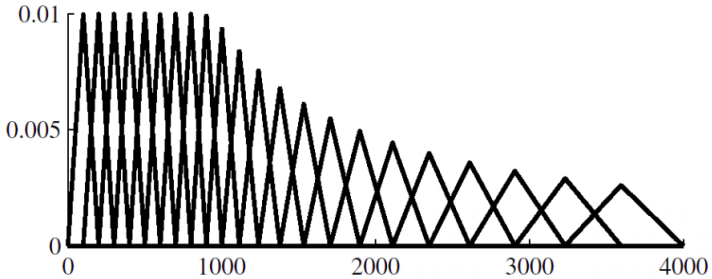


Σχήμα 2.4: *Γραφική απεικόνιση της κλίμακας mel ως προς τη κλίμακα σε Hz.*

κατάσταση [14]. Σε αυτές τις περιπτώσεις, φαίνεται ότι διαφοροποιείται σημαντικά το εύρος ζώνης της συχνότητας φωνοσυντονισμού.

Mel-Frequency Cepstral Coefficients (MFCCs)

Ιδιαίτερα σημαντικά χαρακτηριστικά του σήματος είναι τα MFCCs (Mel-Frequency Cepstral Coefficients), τα οποία κωδικοποιούν και αυτά πληροφορία σχετικά με τη μορφή της φωνητικής οδού. Βασικό στοιχείο τους είναι ότι παρέχουν μια καλύτερη αναπαράσταση του σήματος, αξιοποιώντας ιδιότητες της ακοής του ανθρώπου. Για το σκοπό αυτό, πραγματοποιείται ανάλυση συχνοτήτων με βάση συστοιχία τριγωνικών φίλτρων, των οποίων οι κεντρικές συχνότητες είναι ισοκατανεμημένες στην κλίμακα mel. Τα εύρη ζώνης των φίλτρων είναι σταθερά για κεντρικές συχνότητες μικρότερες από 1 kHz και αυξάνονται εκθετικά μέχρι το μισό της συχνότητας δειγματοληψίας. Στο Σχήμα 2.5, απεικονίζεται η συστοιχία φίλτρων για την περίπτωση σήματος με συχνότητα δειγματοληψίας 8 kHz, σύμφωνα με το [10].



Σχήμα 2.5: *Συστοιχία φίλτρων σε κλίμακα mel για την ανάλυση συχνοτήτων κατά τον υπολογισμό των MFCCs. (Η γραφική απεικόνιση προέρχεται από το [10].)*

Στις περισσότερες υλοποιήσεις, η εξαγωγή των συνιστωσών MFCC απαιτεί, αρχικά, τον υπολογισμό του διακριτού μετασχηματισμού Fourier (DFT) για κάθε πλαίσιο [10]. Ο χωρισμός κάθε εκφώνησης σε πλαίσια έχει πραγματοποιηθεί μετά από πιθανή προ-επεξεργασία του σήματος φωνής και με εφαρμογή παραθύρου, όπως Hamming. Ο μετασχηματισμός DFT για το m-οστό πλαίσιο $x_m[n]$ με N δείγματα ορίζεται ως:

$$X_m[k] = \sum_{n=0}^{N-1} x_m[n] e^{-j(2\pi k/N)n}, \quad k = 0, 1, ..., N-1 \tag{2.3.2}$$

Οι τιμές του DFT, που προκύπτουν για το συγκεκριμένο πλαίσιο, χωρίζονται με χρήση της συστοιχίας τριγωνικών φίλτρων, όπως αναλύθηκε παραπάνω, και σταθμίζονται ανάλογα (βλ. Σχήμα 2.5). Με χρήση R φίλτρων, το mel-φάσμα του πλαισίου προκύπτει:

$$MF_m[r] = \frac{\sum_{k=L_r}^{U_r} |V_r[k] X_m[k]|^2}{\sum_{k=L_r}^{U_r} |V_r[k]|^2}, \quad r = 1, 2, ..., R \tag{2.3.3}$$

όπου $V_r[k]$ η συνάρτηση στάθμισης του r-οστού φίλτρου, το οποίο έχει εύρος από το L_r έως το U_r. Το μέτρο των εξόδων των φίλτρων χρησιμοποιείται για τον υπολογισμό διακριτού μετασχηματισμού συνημιτόνου (DCT), σύμφωνα με τη σχέση:

$$MFCC_m[n] = \frac{1}{R} \sum_{r=1}^{R} log(MF_m[r]) cos\Big[\frac{2\pi}{R}\Big(r + \frac{1}{2}\Big)n\Big], \quad n = 1, 2, ..., N_{MFCC} \tag{2.3.4}$$

όπου ο αριθμός N_{MFCC} των συνιστωσών MFCC είναι μικρότερος από τον αριθμό των φίλτρων R. Για παράδειγμα, τυπικές τιμές είναι $R = 24$ και $N_{MFCC} = 13$.

Ενέργεια βραχέος χρόνου

Πληροφορία σχετικά με τη συναισθηματική κατάσταση του ομιλητή δίνει και η Ενέργεια βραχέος χρόνου, καθώς σχετίζεται με το επίπεδο έντασης των συναισθημάτων [14]. Υπολογίζεται για κάθε πλαίσιο $x_m[n]$ σύμφωνα με τη σχέση:

$$E_m = \frac{1}{N} \sum_{n=0}^{N-1} (x[n])^2 \tag{2.3.5}$$

Ρυθμός Διέλευσης από το Μηδέν

Σε ένα σήμα διακριτού χρόνου, παρατηρείται διέλευση από το μηδέν, όταν δύο διαδοχικά δείγματα του σήματος έχουν διαφορετικό αλγεβρικό πρόσημο [10]. Ο ρυθμός διέλευσης από το μηδέν αντιστοιχεί στο μέσο αριθμό των διελεύσεων από το μηδέν σε κάποια μονάδα χρόνου.

Φασματική Ροή (Spectral Flux)

Η Φασματική Ροή (Spectral Flux) αποτελεί φασματικό χαρακτηριστικό και ορίζεται ως το μέτρο της μεταβολής του πλάτους του φάσματος ενός πλαισίου σε σχέση με το προηγούμενο

πλαίσιο [15]. Σε σύγκριση με τη μουσική, η ομιλία παρουσιάζει μικρότερη φασματική ροή [15]. Στην ομιλία παρατηρούνται τόσο μεταβατικές περίοδοι, για παράδειγμα από ένα φωνήεν σε ένα σύμφωνο, όσο και περίοδοι σταθερότητας, όπως κατά τη διάρκεια ενός φωνήεντος. Αντίθετα, η μουσική παρουσιάζει μεγαλύτερο και πιο σταθερό ρυθμό αλλαγών του φάσματος.

Ποιότητα Φωνής

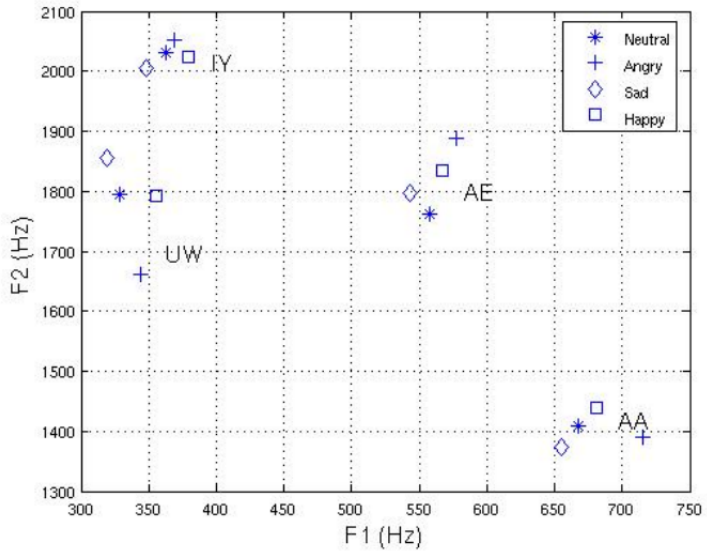
Ενδιαφέρον παρουσιάζουν, επίσης, οι παράμετροι ποιότητας φωνής jitter και shimmer, οι οποίες εκτιμούν τη μέση μεταβολή ανά περίοδο της θεμελιώδους συχνότητας και του πλάτους αντίστοιχα ενός σήματος φωνής [16]. Οι συγκεκριμένες παράμετροι μπορούν να δείξουν, για παράδειγμα, ποιότητες μη φυσιολογικής φωνής, όπως η βραχνή, η ξεψυχισμένη ή η τραχιά φωνή. Επιπλέον, μέτρο ποιότητας φωνής αποτελεί και το Harmonics-to-Noise Ratio (HNR), το οποίο εκφράζει το ποσοστό του πρόσθετου θορύβου στο σήμα φωνής σε dB [17]. Ο θόρυβος αυτός δημιουργείται από το στροβιλισμό του αέρα, κατά την ροή του στη φωνητική οδό. Ο στροβιλισμός αυτός αυξάνεται με την αύξηση της ροής του αέρα, αν οι φωνητικές χορδές δεν κλείνουν αρκετά κατά την παραγωγή κάποιου φωνήματος. Επίσης, θόρυβο προκαλεί και η μη περιοδική ταλάντωση των φωνητικών χορδών. Έτσι, το HNR δείχνει το ποσοστό της περιοδικότητας προς τη μη περιοδικότητα του σήματος φωνής.

Στατιστικά Χαρακτηριστικά

Σε όλα τα παραπάνω χαρακτηριστικά, είναι δυνατή η επιβολή στατιστικών, όπως αναφέρθηκε και στην εισαγωγή [13]. Συνηθισμένα τέτοια στατιστικά είναι: Ακραίες τιμές (Μέγιστη και Ελάχιστη τιμή), Μέσες τιμές (Αριθμητικός ή Γεωμετρικός μέσος όρος), Διάμεσος, Ροπές (Τυπική Απόκλιση, Διακύμανση, Κύρτωση, Λοξότητα), Εκατοστημόρια, Τεταρτημόρια, Κεντροειδή, Κλίση, Μέση τιμή τετραγωνικού σφάλματος, Διάρκεια/Χρόνος. Με τη χρήση αυτών των στατιστικών, προκύπτουν χαρακτηριστικά ανά ομάδα πλαισίων ή ανά εκφώνηση.

2.4 Σχέση με Συναισθηματικές Καταστάσεις

Όπως περιγράφτηκε στο Κεφάλαιο 1, το συναίσθημα επηρεάζει τη φωνή του ανθρώπου, με τέτοιο τρόπο ώστε να γίνεται αντιληπτό από τον συνομιλητή. Όμως, ο άνθρωπος έχει αναπτυγμένο ένα πολύπλοκο σύστημα ακοής και αντίληψης. Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι έρευνες σχετικά με την επιρροή του συναισθήματος στα ακουστικά χαρακτηριστικά του σήματος φωνής, καθώς και στην άρθρωση του ανθρώπου. Γενικά, έχουν παρατηρηθεί διάφορες μεταβολές σχετικά με τη θέση και την κίνηση της γλώσσας, του σαγονιού και των χειλιών, ανάλογα με το συναίσθημα που εκφράζεται [18]. Ενδεικτικά, εδώ, παρατίθεται ένα διάγραμμα απεικόνισης των δύο πρώτων συχνοτήτων φωνοσυντονισμού, κατά την άρθρωση φωνημάτων, όπως προέκυψε στο [18] (βλ. Σχήμα 2.6). Παρατηρείται μεταβολή των τιμών των συχνοτήτων κατά την έκφραση κάθε συναισθήματος, καθώς και διαφορετική επιρροή ανάλογα με το φώνημα.



Σχήμα 2.6: *Μέσες τιμές των F1 και F2 κατά την άρθρωση 4 φωνημάτων με έκφραση ενός συναισθήματος από τα: ουδέτερο, θυμός, λύπη, χαρά. (Η γραφική απεικόνιση προέρχεται από το [18].)*

Πιο συγκεντρωτικά, στον Πίνακα 2.2, αναφέρεται η επιρροή τριών συναισθηματικών καταστάσεων (θυμός, χαρά, λύπη) σε κάποια βασικά ακουστικά χαρακτηριστικά, σύμφωνα με το [1]. Στο συγκεκριμένο πίνακα, παρατηρούμε και προφανή στοιχεία, όπως η υψηλή ένταση φωνής όταν ο ομιλητής είναι θυμωμένος. Η παρατήρηση τέτοιων μεταβολών συμβάλλει στην επιλογή χαρακτηριστικών, τα οποία θα οδηγήσουν σε ορθότερη αναγνώριση του συναισθήματος.

	Θυμός	*Χαρά*	*Λύπη*
Θεμελιώδης Συχνότητα	Αύξηση της μέσης τιμής, της διαμέσου, του εύρους τιμών	Αύξηση της μέσης τιμής, της διαμέσου, του εύρους τιμών	Κάτω από την κανονική μέση τιμή και εύρος τιμών
Ένταση	Υψηλή	Αύξηση	Μείωση
Ενέργεια υψηλών συχνοτήτων		Αύξηση	Μείωση
Ρυθμός ομιλίας	Υψηλός	Αύξηση	Ελαφρά αργός
Ποιότητα Φωνής	τεταμένη, ξεψυχισμένη, πολύ δυνατή	τεταμένη, ξεψυχισμένη , πολύ δυνατή	χαλαρή, βαθιά

Πίνακας 2.2: *Ακουστικά Χαρακτηριστικά και Συναίσθημα. (Οι πληροφορίες προέρχονται από το [1].)*

2.5 Διαφοροποίηση μεταξύ των Ομιλητών

Ιδιαίτερο ενδιαφέρον παρουσιάζει η σύγκριση των διαφορετικών ομιλητών και η μελέτη των στοιχείων που τους διαφοροποιούν. Γενικά, τα εξαγόμενα ακουστικά χαρακτηριστικά εμφανίζουν μεταβολές μεταξύ των ομιλητών, οι οποίες μπορούν να επηρεάσουν σημαντικά την απόδοση ενός συστήματος αναγνώρισης συναισθήματος από φωνή, όπως αναφέρθηκε και στο Κεφάλαιο 1. Οι φωνητικές αυτές μεταβολές οφείλονται σε μία ποικιλία στοιχείων κάθε ατόμου, περιλαμβάνοντας τόσο βιολογικά, όσο και κοινωνικο-πολιτισμικά χαρακτηριστικά. Επιπλέον, ως τέτοιο στοιχείο μπορεί να θεωρηθεί και το περιβάλλον ηχογράφησης κάθε ομιλητή, καθώς επιφέρει μεταβολές στο τελικό σήμα φωνής. Παρακάτω, αναλύονται τα κυριότερα στοιχεία, τα οποία επηρεάζουν την ομιλία ενός ανθρώπου.

2.5.1 Βιολογικά Στοιχεία

Ένα βασικό στοιχείο διαφοροποίησης των ομιλητών αποτελεί το φύλο. Το φύλο επηρεάζει σημαντικά το μήκος της φωνητικής οδού, όπως και τη μορφή των φωνητικών χορδών. Ενδεικτικά, η μέση θεμελιώδης συχνότητα (pitch) ενός άνδρα ομιλητή είναι περίπου 125 Hz, ενώ η αντίστοιχη για τις γυναίκες είναι σχεδόν διπλάσια (περίπου 227 Hz) [10]. Αντίστοιχα, οι συχνότητες φωνοσυντονισμού είναι υψηλότερες στις γυναίκες. Επίσης, σημαντικός παράγοντας είναι και η ηλικία, η οποία καθορίζει τις διαστάσεις της φωνητικής οδού, καθώς και τη λειτουργικότητα των φωνητικών χορδών. Συγκρίνοντας με παραπάνω, η μέση θεμελιώδης συχνότητα των παιδιών βρίσκεται περίπου στα 303 Hz, ακόμα υψηλότερη δηλαδή από τις γυναίκες.

2.5.2 Κοινωνικο-πολιτισμικά Στοιχεία

Στο [19], αναφέρεται η μελέτη των ομοιοτήτων και των διαφορών, μεταξύ Κινέζων και Βόρειων Αμερικάνων, κατά την αναγνώριση 6 βασικών συναισθημάτων. Ως υλικό παρουσίασης χρησιμοποιήθηκε η έκφραση κάθε συναισθήματος από ηθοποιούς Κινέζους και Αμερικάνους, σε 3 πιθανές εκδοχές ερεθίσματος: μόνο ακουστικό, μόνο οπτικό και ο συνδυασμός οπτικο-ακουστικού. Όσον αφορά το καθαρά ακουστικό ερέθισμα, βρέθηκε ότι τα άτομα δυσκολεύονται να αναγνωρίσουν το συναίσθημα όταν εκφράζεται από ομιλητή διαφορετικής κουλτούρας. Αντίθετα, χρησιμοποιούν κυρίως την οπτική πληροφορία. Παρατηρείται, λοιπόν, διαφοροποίηση της έκφρασης και της αντίληψης των συναισθημάτων μεταξύ των πολιτισμών.

Σε συνέχεια των πολιτισμικών διαφορών, αξίζει να αναφερθεί η μελέτη του J. Russell [20] όσον αφορά την κατηγοριοποίηση των συναισθημάτων στις διαφορετικές κουλτούρες. Γενικά, άνθρωποι από διαφορετικούς πολιτισμούς που μιλούν διαφορετικές γλώσσες φαίνεται να διαχωρίζουν τα συναισθήματα με διαφορετικό τρόπο. Επίσης, οι λέξεις που περιγράφουν συναισθήματα, όπως θυμός ή χαρά, δεν είναι κοινές σε όλες τις γλώσσες, όπως δεν είναι και κοινά τα όρια μεταξύ διαφορετικών συναισθημάτων. Επιπλέον, ιδιαίτερη επιρροή κάθε κουλτούρας παρατηρείται στη συναισθηματική έκφραση, καθώς και στη συμπεριφορά των ανθρώπων [21]. Το κοινωνικό περιβάλλον επιβάλλει αντιλήψεις και κανονισμούς, διαμορφώνοντας έτσι μια κοινή αποδεκτή συμπεριφορά και έκφραση μεταξύ των ανθρώπων. Παρόμοια επιρροή έχει και η θρησκεία.

Η έκφραση ενός συναισθήματος προκύπτει ότι εξαρτάται και από τη γλώσσα του ομιλητή [22]. Θεωρώντας ότι σημαντικά προσωδιακά χαρακτηριστικά του σήματος φωνής, που επηρεάζο-

νται από το συναίσθημα, είναι η θεμελιώδης συχνότητα, η διάρκεια και η ενέργεια των συλλαβών και γνωρίζοντας ότι αυτά διαφοροποιούνται μεταξύ των γλωσσών, φαίνεται ότι το συναίσθημα είναι εξαρτώμενο της γλώσσας σε κάποιο βαθμό. Ενδεικτικά, στην αγγλική γλώσσα, μια τονισμένη συλλαβή προφέρεται με μεγαλύτερη ένταση, θεμελιώδη συχνότητα και διάρκεια από μία μη τονισμένη. Αντίθετα, στην ιαπωνική γλώσσα το μόνο σημαντικό στοιχείο που καθορίζει την προφορά είναι η θεμελιώδης συχνότητα.

Κλείνοντας τα κοινωνικο-πολιτισμικά στοιχεία που διαφοροποιούν τους ομιλητές, είναι απαραίτητο να αναφερθεί και ο προσωπικός χαρακτήρας και η συμπεριφορά του κάθε ατόμου. Άλλος θυμώνει πιο εύκολα και με μεγαλύτερη ένταση, άλλος είναι χαμηλών τόνων, άλλος ενθουσιάζεται με κάτι, ενώ άλλος εκφράζει πιο ήπια τη χαρά του. Όμως, ένα τηλεφωνικό κέντρο ή ένα σύστημα αυτόματης διδασκαλίας είναι πολύ δύσκολο να διακρίνει μικρές τέτοιες διαφορές και να προσαρμόσει κατάλληλα την αντίδρασή του. Για παράδειγμα, μπορεί να μη γίνει αντιληπτός ο θυμός ενός ατόμου, αν είναι πιο μικρής έντασης από τις θυμωμένες φράσεις των ομιλητών, με τους οποίους το σύστημα έχει εκπαιδευτεί.

Κεφάλαιο 3

Ταξινομητές

3.1 Εισαγωγή

Στο φυσικό κόσμο, είναι ανεπτυγμένα πολλά συστήματα αναγνώρισης προτύπων. Για παράδειγμα, ο άνθρωπος έχει την ικανότητα να αναγνωρίζει ένα πρόσωπο ή έναν ήχο, με βάση κάποια χαρακτηριστικά. Αντίστοιχα, αντιλαμβάνεται εύκολα αν ο συνομιλητής του είναι θυμωμένος ή χαρούμενος και μπορεί να προσαρμόζει ανάλογα την επικοινωνία του. Η ικανότητα αυτή προέρχεται από διαρκή εκπαίδευση του νευρικού του συστήματος στο περιβάλλον του. Κάθε τέτοια περίπτωση μπορεί να θεωρηθεί ως λήψη απόφασης, με την έννοια ότι υπάρχει ένα σύνολο επιλογών, οι οποίες είναι γνωστές. Όμως, η υιοθέτηση οποιασδήποτε απόφασης προϋποθέτει κάποια έννοια ταξινόμησης των δυνατών επιλογών.

Στον τομέα της Αναγνώρισης Προτύπων, έχουν αναπτυχθεί διάφοροι αλγόριθμοι ταξινόμησης, με σκοπό την ανάπτυξη ικανότητας λήψης απόφασης από τις μηχανές. Τέτοιοι αλγόριθμοι είναι απαραίτητοι σε διάφορες επιστημονικές περιοχές, όπως η Αναγνώριση Φωνής, η Όραση Υπολογιστών και η Αναγνώριση Συναισθήματος. Σε κάθε μία από αυτές τις περιοχές, υπάρχει πληθώρα προβλημάτων ταξινόμησης, με σκοπό την τεχνολογική ανάπτυξη. Συνήθως, ένα πρόβλημα ταξινόμησης περιλαμβάνει έναν αριθμό κλάσεων $c_1, c_2, ..., c_C$, στις οποίες είναι δυνατόν να κατηγοριοποιηθεί οποιοδήποτε δεδομένου εισόδου. Μετά την εξαγωγή των χαρακτηριστικών των δεδομένων, οι αλγόριθμοι ταξινόμησης αναπτύσσουν μοντέλα, με σκοπό την ταξινόμηση των δεδομένων αυτών, με βάση τα χαρακτηριστικά τους.

Για την ανάπτυξη κάθε τέτοιου μοντέλου ή ταξινομητή, είναι απαραίτητη αρχικά η εκπαίδευσή του. Για το σκοπό αυτό, χρησιμοποιούνται τα δεδομένα εκπαίδευσης. Στη συνέχεια, η απόδοσή του ελέγχεται με βάση τα δεδομένα αξιολόγησης. Όσον αφορά την εκπαίδευση ενός ταξινομητή, υπάρχουν δύο βασικοί τρόποι μάθησης: με επίβλεψη (*supervised*) και χωρίς επίβλεψη (*unsupervised*). Στην πρώτη περίπτωση, είναι διαθέσιμα επισημειωμένα δεδομένα, με σκοπό την ανάπτυξη του μοντέλου. Αντίθετα, στη μάθηση χωρίς επίβλεψη δεν είναι διαθέσιμες οι ετικέτες των δεδομένων εκπαίδευσης, και άρα δεν υπάρχει πληροφορία σχετικά με την κλάση, στην οποία ανήκει το καθένα από αυτά.

Αναφορικά με τον τομέα της Αναγνώρισης Συναισθήματος από Φωνή, έχουν χρησιμοποιηθεί διάφοροι ταξινομητές, τόσο στην κλασσική μορφή τους, όσο και με ένταξή τους σε πιο πολύπλοκα μοντέλα. Οι πιο βασικοί από αυτούς περιλαμβάνουν: Μοντέλα Μείγματος Gaussian Συνιστωσών (GMM) [23, 24], Support Vector Machine (SVM) [23, 24, 25, 26, 27] και

Νευρωνικά Δίκτυα [23, 28, 29]. Επίσης, συναντάται η προσαρμογή ενός βασικού GMM σε κάθε δεδομένο και η εξαγωγή του αντίστοιχου GMM υπερ-διανύσματος (διάνυσμα ορισμένο ως η αλληλουχία των μέσων τιμών των Gaussian συνιστωσών του μείγματος), με σκοπό την ταξινόμησή τους με SVM [30, 24].

Στο παρόν κεφάλαιο, αναλύονται τα μοντέλα που θα χρησιμοποιηθούν για ταξινόμηση σε επόμενα κεφάλαια: Μοντέλα Μείγματος Gaussian Συνιστωσών (παρ. 3.3) και Support Vector Machine (παρ. 3.4). Επίσης, γίνεται αναφορά στη Θεωρία Αποφάσεων κατά Bayes, καθώς αποτελεί βάση εννοιών που θα ακολουθήσουν. Τέλος, δίνεται μια συνοπτική περιγραφή του *Affective Saliency Model* [31], το οποίο θα αξιοποιηθεί στο Κεφάλαιο 4.

3.2 Ταξινόμηση κατά Bayes

3.2.1 Θεωρία Αποφάσεων κατά Bayes

Η Θεωρία Αποφάσεων κατά Bayes αποτελεί βασική στατιστική προσέγγιση της Αναγνώρισης Προτύπων [32], βασιζόμενη σε έννοιες της θεωρίας των πιθανοτήτων. Ποσοτικοποιεί τις αποφάσεις ενός προβλήματος ταξινόμησης, χρησιμοποιώντας πιθανότητες και κόστη. Τελικά, λόγω της στατιστικής φύσης των προτύπων, ταξινομεί ένα άγνωστο πρότυπο στην πιο πιθανή κλάση [33].

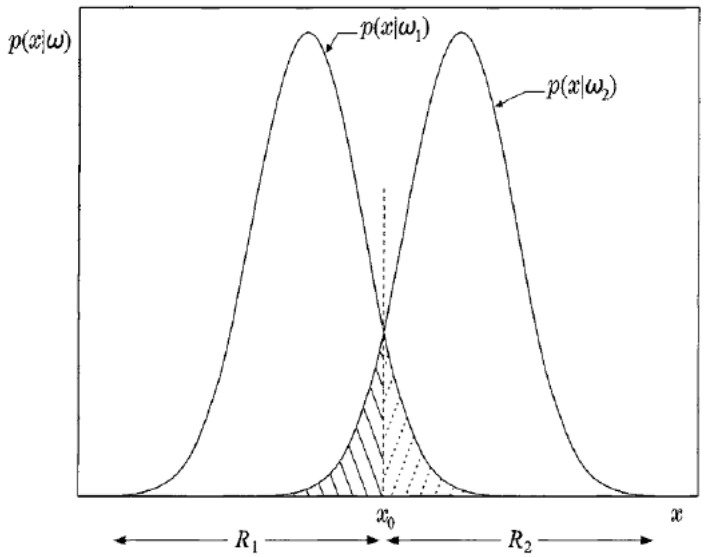
Έστω ένα πρόβλημα ταξινόμησης που περιλαμβάνει τις κλάσεις $c_1, c_2, ..., c_C$ και το διάνυσμα χαρακτηριστικών $\mathbf{x}$ ενός άγνωστου προτύπου. Τότε ο *κανόνας του Bayes* δίνει την εκ-των-υστέρων (a-posteriori) πιθανότητα της κλάσης c_i, σύμφωνα με τη σχέση:

$$P(c_i|\mathbf{x}) = \frac{p(\mathbf{x}|c_i)P(c_i)}{p(\mathbf{x})} \tag{3.2.1}$$

Δηλαδή, η εκ-των-υστέρων πιθανότητα $P(c_i|\mathbf{x})$ ισούται με το γινόμενο της συνάρτησης πιθανοφάνειας $p(\mathbf{x}|c_i)$ της κλάσης c_i επί την εκ-των-προτέρων πιθανότητα $P(c_i)$, προς την συνάρτηση πυκνότητας πιθανότητας $p(\mathbf{x})$. Οι εκ-των-προτέρων (a-priori) πιθανότητες των κλάσεων θεωρούνται γνωστές, αντλώντας πληροφορία από τα δεδομένα εκπαίδευσης. Με βάση την παραπάνω έκφραση, ο κανόνας ταξινόμησης κατά Bayes ταξινομεί το άγνωστο πρότυπο με διάνυσμα χαρακτηριστικών $\mathbf{x}$ στην κλάση c_i, εφόσον ισχύει:

$$P(c_i|\mathbf{x}) > P(c_j|\mathbf{x}), \quad \forall j \neq i \tag{3.2.2}$$

Αν θεωρηθούν μόνο 2 κλάσεις και η περίπτωση ενός χαρακτηριστικού x, τότε σύμφωνα με τα παραπάνω, δημιουργείται ένα σημείο απόφασης x_0, όπως απεικονίζεται στο Σχήμα 3.1. Το σημείο απόφασης αυτό χωρίζει τον άξονα x σε 2 περιοχές R_1 και R_2. Αν ένα άγνωστο πρότυπο αντιπροσωπεύεται από τιμή του x στην περιοχή R_1, τότε ταξινομείται στην κλάση c_1. Αντίστοιχα ισχύει για την περιοχή R_2. Υπάρχει, όμως, περίπτωση το πρότυπο με $x \in R_1$, να ανήκει στην κλάση c_2, και αντίστροφα. Τότε, η ταξινόμηση κρίνεται εσφαλμένη. Η πιθανότητα του σφάλματος ταξινόμησης αντιστοιχεί στο γραμμοσκιασμένο εμβαδόν του σχήματος. Αποδεικνύεται ότι η επιλογή του ταξινομητή κατά Bayes ελαχιστοποιεί το σφάλμα ταξινόμησης [33].



Σχήμα 3.1: *Παράδειγμα σημείου απόφασης, σύμφωνα με τον κανόνα ταξινόμησης κατά Bayes, για την περίπτωση 2 κλάσεων* ω_1, ω_2.

3.2.2 Εκτίμηση Μέγιστης Πιθανοφάνειας

Στην προηγούμενη προσέγγιση, θεωρούνται γνωστές οι συναρτήσεις πυκνότητας πιθανότητας. Συχνά, όμως, ένα πρόβλημα ταξινόμησης απαιτεί την εκτίμηση των συναρτήσεων αυτών και των παραμέτρων τους, με βάση τα διαθέσιμα δεδομένα. Έχουν αναπτυχθεί, λοιπόν, διάφοροι τρόποι προσέγγισης αυτού του προβλήματος εκτίμησης των παραμέτρων. Εδώ, θα αναφερθεί μια βασική μέθοδος, που ονομάζεται Μέθοδος της Μέγιστης Πιθανοφάνειας (Maximum Likelihood). Η συγκεκριμένη μέθοδος είναι αρκετά απλή συγκριτικά με εναλλακτικές μεθόδους, όπως επίσης εμφανίζει καλές ιδιότητες σύγκλισης με την αύξηση του αριθμού των δεδομένων εκπαίδευσης [32].

Αρχικά, υποθέτει ότι οι συναρτήσεις πυκνότητας πιθανότητας $p(\mathbf{x}|c_i;\boldsymbol{\vartheta}_i)$ δίνονται σε παραμετρική μορφή και το ζητούμενο είναι η εκτίμηση των παραμέτρων τους $\boldsymbol{\vartheta}_i$. Επιπλέον, για απλοποίηση του προβλήματος θεωρείται ότι τα δεδομένα μίας κλάσης δε δίνουν πληροφορία για τις άλλες κλάσεις, με σκοπό τη μελέτη κάθε κλάσης ξεχωριστά. Έτσι, θεωρώντας ένα σύνολο τυχαίων και ανεξάρτητων δειγμάτων $D = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ και την από κοινού συνάρτηση πυκνότητας πιθανότητας $p(D|\boldsymbol{\vartheta})$, η Μέγιστη Πιθανοφάνεια εκτιμά το $\boldsymbol{\vartheta}$ σύμφωνα με τη σχέση:

$$\hat{\boldsymbol{\vartheta}}_{ML} = \arg\max_{\boldsymbol{\vartheta}} \mathcal{L}(\boldsymbol{\vartheta}) = \arg\max_{\boldsymbol{\vartheta}} p(D|\boldsymbol{\vartheta}) = \arg\max_{\boldsymbol{\vartheta}} \prod_{k=1}^{N} p(\mathbf{x}_k|\boldsymbol{\vartheta}) \tag{3.2.3}$$

όπου $\mathcal{L}(\boldsymbol{\vartheta})$ ονομάζεται συνάρτηση πιθανοφάνειας του $\boldsymbol{\vartheta}$.

3.2.3 Απλοϊκός Ταξινομητής κατά Bayes

Γενικά, η εκτίμηση των παραμέτρων των συναρτήσεων πυκνότητας πιθανότητας $p(\mathbf{x}|c_i)$, οι οποίες αξιοποιούνται από τον κανόνα ταξινόμησης του Bayes, απαιτεί ένα μεγάλο αριθμό δεδομένων εκπαίδευσης για να είναι καλή. Μάλιστα, η ανάγκη αυτή αυξάνει εκθετικά με τη διάσταση d του

χώρου των χαρακτηριστικών [33]. Συχνά, λόγω του περιορισμένου αριθμού των δεδομένων, υιοθετούνται απλοποιήσεις και παραχωρήσεις σχετικά με το βαθμό ακρίβειας των εκτιμήσεων. Μια τέτοια απλοποιημένη προσέγγιση αφορά τα επιμέρους χαρακτηριστικά ενός δείγματος, τα οποία μπορούν να θεωρηθούν στατιστικώς ανεξάρτητα. Σε αυτή την περίπτωση, ισχύει:

$$p(\mathbf{x}|c_i) = \prod_{j=1}^{d} p(x_j|c_i) \tag{3.2.4}$$

Με την παραπάνω προσέγγιση, τελικά απαιτείται αριθμός δεδομένων πολλαπλάσιος της διάστασης d, αντί της προηγούμενης εκθετικής σχέσης. Υιοθετώντας τη θεώρηση αυτή, ο Απλοϊκός Ταξινομητής κατά Bayes (Naive Bayes) ταξινομεί ένα άγνωστο δείγμα στην κλάση:

$$\hat{c} = \arg\max_{c_i} P(c_i) \prod_{j=1}^{d} p(x_j|c_i) \tag{3.2.5}$$

Στο [33] αναφέρεται ότι ο Απλοϊκός Bayes Ταξινομητής μπορεί να είναι αρκετά εύρωστος σε περιπτώσεις παραβίασης της υπόθεσης ανεξαρτησίας. Όσον αφορά τις συναρτήσεις πυκνότητας πιθανότητας, μια συχνή επιλογή κατανομής είναι η Gaussian.

3.3 Μοντέλο Μείγματος Gaussian Συνιστωσών (GMM)

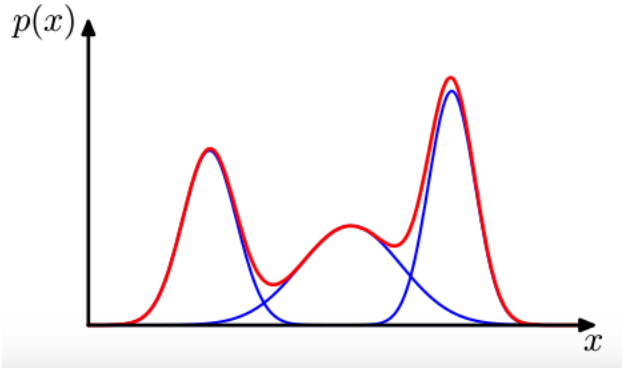
Ένα τρόπος μοντελοποίησης μιας άγνωστης συνάρτησης πυκνότητας πιθανότητας, ευρέως διαδεδομένος, βασίζεται σε Μοντέλο Μείγματος Gaussian Συνιστωσών (Gaussian Mixture Model ή GMM). Ουσιαστικά, θεωρείται ότι η συνάρτηση αυτή ισούται με το γραμμικό συνδυασμό M Gaussian κατανομών, σύμφωνα με τη σχέση:

$$p(\mathbf{x}|\boldsymbol{\vartheta}) = \sum_{m=1}^{M} w_m \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{3.3.1}$$

όπου $\mathbf{x}$ το διάνυσμα των χαρακτηριστικών, w_m το βάρος που αντιστοιχεί στη συνιστώσα m, για $m = 1, ..., M$, και $\boldsymbol{\vartheta}$ το διάνυσμα των παραμέτρων του μοντέλου που περιλαμβάνουν το διάνυσμα μέσης τιμής $\boldsymbol{\mu}_m$, τον πίνακα συνδιακύμανσης $\boldsymbol{\Sigma}_m$ και το βάρος κάθε Gaussian συνιστώσας. Ισχύει $\sum_{m=1}^{M} w_m = 1$. Η γενική μορφή μιας Gaussian συνιστώσας d διαστάσεων δίνεται από τη σχέση:

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\Big[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big] \tag{3.3.2}$$

Συχνά, χρησιμοποιείται διαγώνιος πίνακας συνδιακύμανσης. Στο Σχήμα 3.2, απεικονίζεται ένα παράδειγμα Μείγματος Gaussian Συνιστωσών για την περίπτωση ενός μόνο χαρακτηριστικού ($d = 1$) και τριών συνιστωσών ($M = 3$).



Σχήμα 3.2: *Παράδειγμα Μείγματος Gaussian Συνιστωσών για την περίπτωση ενός μόνο χαρακτηριστικού. (Η γραφική απεικόνιση προέρχεται από το [34].)*

Το ζητούμενο της παραπάνω μοντελοποίησης είναι η εκτίμηση των παραμέτρων $\boldsymbol{\vartheta}$. Η εκτίμηση αυτή είναι δυνατόν να γίνει με εφαρμογή του Expectation Maximization αλγορίθμου [34], ο οποίος περιλαμβάνει τα δύο βήματα που αναφέρονται ακολούθως. Θεωρώντας, αρχικά, ένα σύνολο δειγμάτων $D = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, υποθέτεται ότι είναι ανεξάρτητα μεταξύ τους και αρχικοποιούνται οι παράμετροι $\boldsymbol{\vartheta}$ του μοντέλου. Το πρώτο βήμα (*Βήμα E*) αφορά την εκτίμηση των εκ-των-υστέρων πιθανοτήτων, σύμφωνα με τη σχέση:

$$\gamma_{km} = \frac{w_m \mathcal{N}(\mathbf{x}_k; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{j=1}^{M} w_j \mathcal{N}(\mathbf{x}_k; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{3.3.3}$$

Ακολουθεί το βήμα Μεγιστοποίησης (*Βήμα M*), όπου προκύπτουν οι παρακάτω σχέσεις εκτίμησης των νέων τιμών των παραμέτρων:

$$\boldsymbol{\mu}'_m = \frac{\sum_{k=1}^{N} \gamma_{km} \mathbf{x}_k}{\sum_{k=1}^{N} \gamma_{km}} \tag{3.3.4}$$

$$\boldsymbol{\Sigma}'_m = \frac{\sum_{k=1}^{N} \gamma_{km} (\mathbf{x}_k - \boldsymbol{\mu}'_m)(\mathbf{x}_k - \boldsymbol{\mu}'_m)^\intercal}{\sum_{k=1}^{N} \gamma_{km}} \tag{3.3.5}$$

$$w'_m = \frac{1}{N} \sum_{k=1}^{N} \gamma_{km} \tag{3.3.6}$$

Η διαδικασία επαναλαμβάνεται, επιστρέφοντας στο βήμα Ε, μέχρι την ικανοποίηση κάποιου κριτηρίου σύγκλισης. Για τη σύγκλιση ελέγχονται είτε οι τιμές των παραμέτρων, είτε η λογαριθμική συνάρτηση πιθανοφάνειας.

3.4 Support Vector Machine (SVM)

Στις προηγούμενες παραγράφους, η ταξινόμηση βασίστηκε σε θεωρία πιθανοτήτων. Εκτός από αυτή τη λογική, έχουν αναπτυχθεί διάφοροι άλλοι ταξινομητές, μεταξύ των οποίων και οι γραμμικοί. Σε αυτή την κατηγορία ανήκουν τα Support Vector Machines (SVM), τα οποία

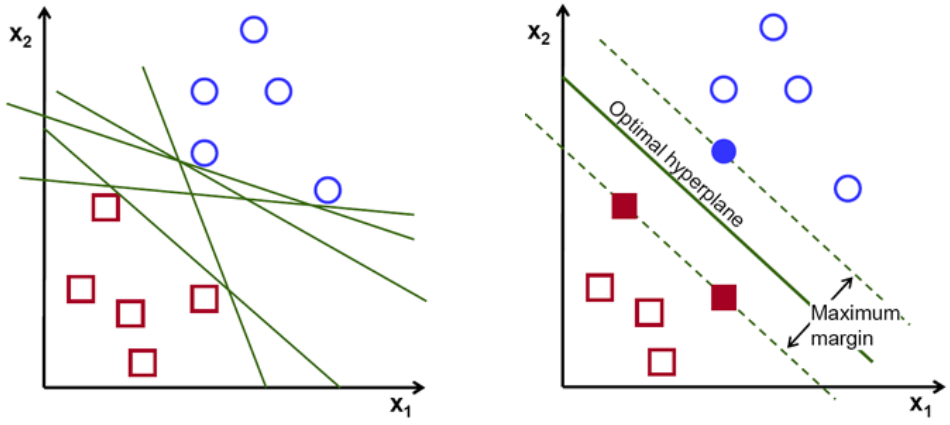
αποτελούν μοντέλα ιδιαίτερα χρήσιμα τόσο στην Αναγνώριση Συναισθήματος από Φωνή, όσο και σε άλλους τομείς. Θεωρώντας 2 κλάσεις, η βασική ιδέα του SVM περιλαμβάνει την εύρεση ενός υπερεπιπέδου που θα τις διαχωρίζει. Σε περίπτωση περισσότερων κλάσεων, έχουν αναπτυχθεί διάφοροι τρόποι χρήσης του βασικού SVM, όπως οι τεχνικές μία-έναντι-των-υπολοίπων ή μία-έναντι-μίας [33], όπου γίνεται σύγκριση ανά δύο σύνολα με βάση κάποια ιεραρχία.

3.4.1 Γραμμικά Διαχωρίσιμες Κλάσεις

Αρχικά, θα γίνει αναφορά στην περίπτωση 2 γραμμικά διαχωρίσιμων κλάσεων c_1 και c_2. Έστω $\mathbf{x}_k$ τα διανύσματα χαρακτηριστικών των δεδομένων εκπαίδευσης, για $k = 1, ..., N$. Στόχος είναι η σχεδίαση ενός υπερεπιπέδου:

$$g(\mathbf{x}) = \mathbf{w}^\intercal \mathbf{x} + w_0 = 0 \tag{3.4.1}$$

με σκοπό την ορθή ταξινόμησή τους. Μελετώντας τα δείγματα των 2 κλάσεων του Σχήματος 3.3, παρατηρούμε ότι είναι δυνατός ο γραμμικός διαχωρισμός τους με παραπάνω από ένα υπερεπίπεδο. Ο ταξινομητής SVM βρίσκει το βέλτιστο υπερεπίπεδο, μεγιστοποιώντας το *περιθώριο* (*margin*) μεταξύ του υπερεπιπέδου και της κάθε κλάσης, όπως απεικονίζεται στο δεξιά σχήμα. Ουσιαστικά, το βέλτιστο υπερεπίπεδο θα έχει την ίδια απόσταση από τα αντίστοιχα πλησιέστερα δείγματα των 2 κλάσεων.



Σχήμα 3.3: *Παράδειγμα προβλήματος 2 γραμμικά διαχωρίσιμων κλάσεων. Στόχος του SVM είναι η εύρεση του βέλτιστου υπερεπίπεδου, που δίνει το μέγιστο δυνατό περιθώριο (margin). (Η γραφική απεικόνιση προέρχεται από το [35].)*

Γνωρίζοντας ότι η απόσταση ενός σημείου από το υπερεπίπεδο $g(\mathbf{x})$ ισούται με $\frac{|g(\mathbf{x})|}{||\mathbf{w}||}$ και θεωρώντας ότι η τιμή της $g(\mathbf{x})$ στα πλησιέστερα σημεία των κλάσεων ισούται με 1 για την κλάση c_1 και με -1 για τη c_2, προκύπτει το περιθώριο:

$$\frac{1}{||\mathbf{w}||} + \frac{1}{||\mathbf{w}||} = \frac{2}{||\mathbf{w}||} \tag{3.4.2}$$

Με βάση τα παραπάνω, για την εύρεση των παραμέτρων του βέλτιστου υπερεπιπέδου, ζητείται η ελαχιστοποίηση της συνάρτησης:

$$J(\mathbf{w}, w_0) = \frac{1}{2}||\mathbf{w}||^2 \tag{3.4.3}$$

υπό τους περιορισμούς:

$$z_k(\mathbf{w}^\intercal \mathbf{x}_k + w_0) \geq 1, \quad k = 1, ..., N \tag{3.4.4}$$

όπου θεωρείται $z_k = 1$ για την κλάση c_1 και $z_k = -1$ για την κλάση c_2. Η λύση προκύπτει με χρήση πολλαπλασιαστών Lagrange [33].

3.4.2 Μη Γραμμικά Διαχωρίσιμες Κλάσεις

Η περίπτωση των γραμμικά διαχωρίσιμων κλάσεων, που αναφέρεται παραπάνω, δεν παρατηρείται συχνά στην πραγματικότητα. Τις περισσότερες φορές, δεν υπάρχει υπερεπίπεδο που να μπορεί να διαχωρίσει τέλεια τα δείγματα 2 κλάσεων. Αντίθετα, κάθε υπερεπίπεδο θα οδηγεί σε έναν αριθμό λανθασμένα ταξινομημένων δειγμάτων. Σε αυτή την περίπτωση, είναι δυνατή η εισαγωγή μιας μεταβλητής ξ_k, γνωστής ως *μεταβλητής χαλαρότητας* [33]. Στόχος τώρα είναι η μεγιστοποίηση του περιθωρίου όπως πριν, αλλά ταυτόχρονα ο αριθμός των σημείων για τα οποία ισχύει $z_k(\mathbf{w}^\intercal \mathbf{x}_k + w_0) < 1$ να κρατηθεί όσο το δυνατόν πιο μικρός. Τελικά, απαιτείται η ελαχιστοποίηση της συνάρτησης:

$$J(\mathbf{w}, w_0) = \frac{1}{2}||\mathbf{w}||^2 + C\sum_{k=1}^{N}\xi_k \tag{3.4.5}$$

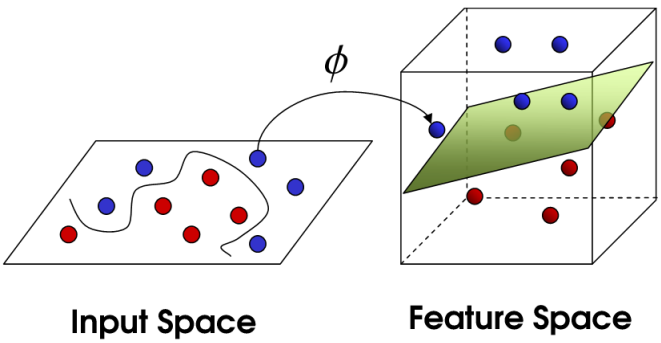
όπου C μία θετική σταθερά, υπό τους περιορισμούς:

$$z_k(\mathbf{w}^\intercal \mathbf{x}_k + w_0) \geq 1 - \xi_k, \quad k = 1, ..., N \tag{3.4.6}$$

$$\xi_k \geq 0, \quad k = 1, ..., N \tag{3.4.7}$$

Η λύση και εδώ προκύπτει με χρήση πολλαπλασιαστών Lagrange [33].

3.4.3 Συναρτήσεις Πυρήνα (Kernels)



Σχήμα 3.4: *Παράδειγμα απεικόνισης των χαρακτηριστικών σε νέο χώρο περισσότερων διαστάσεων, με σκοπό τον διαχωρισμό των 2 κλάσεων. (Η εικόνα προέρχεται από το [36].)*

Στην πράξη, το SVM συχνά απεικονίζει τα χαρακτηριστικά των δειγμάτων εισόδου σε ένα νέο χώρο, συνήθως περισσότερων διαστάσεων. Σκοπός είναι να γίνει εφικτός ο γραμμικός διαχωρισμός των 2 κλάσεων. Με τη βοήθεια κατάλληλης μη γραμμικής απεικόνισης $\varphi()$, αποδεικνύεται

ότι είναι πάντα δυνατός ο διαχωρισμός 2 κλάσεων με χρήση υπερεπιπέδου [32]. Ένα παράδειγμα τέτοιας απεικόνισης, από τον αρχικό χώρο των χαρακτηριστικών στον νέο, ο οποίος ονομάζεται *feature space*, παρουσιάζεται στο Σχήμα 3.4. Με τον τρόπο αυτό, το SVM λειτουργεί, στη συνέχεια, με χρήση των μετασχηματισμένων χαρακτηριστικών $\mathbf{y}_k = \varphi(\mathbf{x}_k)$.

Λόγω της εμφάνισης εσωτερικών γινομένων στη λύση του συγκεκριμένου προβλήματος, αρκεί μόνο ο καθορισμός της συνάρτησης πυρήνα (Kernel), η οποία ορίζεται:

$$K(\mathbf{x}, \mathbf{x}_i) = \varphi(\mathbf{x}_i)^\intercal \varphi(\mathbf{x}) \tag{3.4.8}$$

Στη πράξη χρησιμοποιούνται διάφορες συναρτήσεις ως συναρτήσεις πυρήνα. Κάποιες από τις πιο βασικές είναι η γραμμική, η πολυωνυμική και η RBF (radial basis function).

3.5 Affective Saliency Model

Το *Affective Saliency Model* αποτελεί μοντέλο που αναπτύχθηκε από τους Χωριανοπούλου Α. κ.ά. [31]. Το μοντέλο αυτό συνδυάζει τα εξαγόμενα ακουστικά χαρακτηριστικά με τις εκ-των-υστέρων πιθανότητες ενός ταξινομητή. Σκοπός είναι η εκτίμηση της ποσότητας της συναισθηματικής πληροφορίας στο χρόνο. Βασικό χαρακτηριστικό του αποτελεί η εκτίμηση βαρών με χρήση της τεχνικής Minimum Classification Error (MCE) [37, 38]. Τα βάρη αυτά σταθμίζουν τις εκ-των-υστέρων πιθανότητες ανά ομάδα πλαισίων (segment), με σκοπό τον υπολογισμό της εκ-των-υστέρων πιθανότητας μιας εκφώνησης. Θεωρείται ότι κάθε πλαίσιο, όπως και κάθε ομάδα πλαισίων που ανήκουν σε μία εκφώνηση, αντιστοιχούν στην ίδια συναισθηματική κλάση, σύμφωνα με την ετικέτα της συγκεκριμένης εκφώνησης.

Επιλέχθηκε η χρήση ομάδων πλαισίων, καθώς αποδείχτηκε πιο εύρωστη συγκριτικά με τον υπολογισμό πιθανοτήτων σε επίπεδο πλαισίου. Τα χαρακτηριστικά κάθε ομάδας προκύπτουν με ομαδοποίηση 20 πλαισίων, ξεχωριστά για κάθε εκφώνηση, και υπολογισμό 5 στατιστικών (ελάχιστη τιμή, μέγιστη τιμή, μέση τιμή, διάμεσος και τυπική απόκλιση) κατά μήκος αυτών των πλαισίων. Έστω $\mathbf{x} = [\mathbf{x}_1, ..., \mathbf{x}_N]$ το διάνυσμα των χαρακτηριστικών των ομάδων πλαισίων για μία εκφώνηση και C_i μία κλάση, τότε η εκ-των-υστέρων πιθανότητα της κλάσης αυτής δεδομένου της εκφώνησης προκύπτει:

$$F(C_i|\mathbf{x}) = logP(C_i|\mathbf{x}) = \frac{1}{N}\sum_{j=1}^{N} \lambda_j logP(C_i|\mathbf{x}_j) \tag{3.5.1}$$

Για τον υπολογισμό των βαρών λ_j χρησιμοποιούνται τα *Regression Features* d_k της συγκεκριμένης εκφώνησης. Τα χαρακτηριστικά d_k, στο συγκεκριμένο άρθρο, αποτελούνται από μία ποικιλία χαρακτηριστικών σε επίπεδο πλαισίου και ομάδας πλαισίων, καθώς επίσης και από το ρυθμό άφωνων πλαισίων ανά ομάδα, όπως προκύπτει από το Voice Activity Detector [39]. Με τη χρήση αυτών, εκπαιδεύεται το Regression μοντέλο:

$$\lambda_j = \sum_{k=1}^{K} a_k d_k \tag{3.5.2}$$

όπου a_k είναι τα βάρη εκπαίδευσης, για τα οποία ισχύει $\sum_{k=1}^{K} a_k = 1$. Τα βάρη αυτά εκτιμώνται με χρήση της τεχνικής Minimum Classification Error (MCE), με σκοπό την ελαχιστοποίηση

του ρυθμού λαθών ταξινόμησης, όπου υλοποιείται Generalized Probabilistic Descent (GPD) αλγόριθμος.

Θεωρούνται 2 συναισθηματικές κλάσεις C_1 και C_2. Έστω ότι η εκφώνηση με διάνυσμα $\mathbf{x}$ ανήκει στην κλάση C_1, τότε το λάθος ταξινόμησης ορίζεται ως:

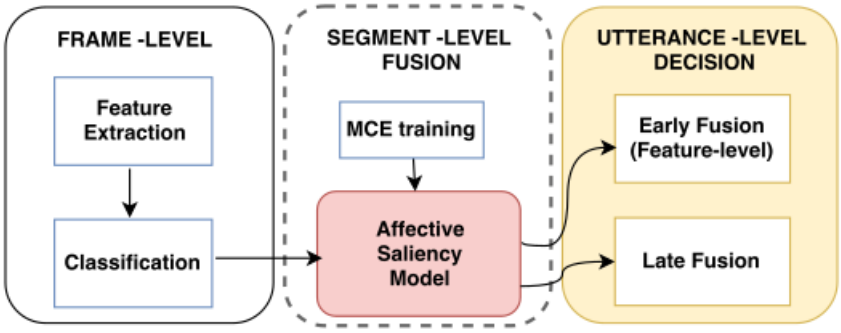
$$E(\mathbf{x}) = F(C_2|\mathbf{x}) - F(C_1|\mathbf{x}) \tag{3.5.3}$$

Στη συνέχεια, το μέτρο E απεικονίζεται στο διάστημα [0,1], μέσω της σιγμοειδής συνάρτησης:

$$l(\mathbf{x}) = \frac{1}{1 + e^{-\gamma E(\mathbf{x})}}, \quad \gamma > 1 \tag{3.5.4}$$

Η παραπάνω συνάρτηση αναφέρεται ως *loss function* και είναι αυτή που παραγωγίζεται, με σκοπό την εύρεση του ελαχίστου, μέσω ενός επαναληπτικού Gradient Descent αλγορίθμου.

Στο συγκεκριμένο άρθρο, υλοποιούνται συνολικά 4 μοντέλα, με στόχο την αξιολόγηση του προτεινόμενου *Affective Saliency Model*: *Baseline*, *Early Fusion*, *Pre-MCE*, *Post-MCE*. Ουσιαστικά γίνεται χρήση ή όχι των βαρών λ_j που προκύπτουν σύμφωνα με την παραπάνω προσέγγιση, σε SVM ή Απλοϊκό Bayes ταξινομητή. Παρακάτω, αναφέρεται μια σύντομη περιγραφή τους. Στο Σχήμα 3.5, απεικονίζεται συγκεντρωτικά η αρχιτεκτονική του συστήματος και των επιλογών του.



Σχήμα 3.5: *Αρχιτεκτονική του συστήματος με χρήση του Affective Saliency Model. (Η εικόνα προέρχεται από το [31].)*

3.5.1 Baseline

Ως μοντέλο αναφοράς (*Baseline*), στο συγκεκριμένο άρθρο χρησιμοποιείται SVM ταξινομητής με γραμμική συνάρτηση πυρήνα. Για την εκπαίδευση και αξιολόγηση του συγκεκριμένου SVM, χρησιμοποιούνται τα ακουστικά χαρακτηριστικά (στατιστικά) σε επίπεδο εκφώνησης. Η υλοποίηση του SVM γίνεται με τη βοήθεια του εργαλείου WEKA [40].

3.5.2 Early Fusion

Ως Early Fusion, αναφέρεται η περίπτωση χρήσης SVM με γραμμική συνάρτηση πυρήνα, όπως παραπάνω, για την ταξινόμηση των σταθμισμένων στατιστικών των αντίστοιχων χαρακτηριστικών ανά εκφώνηση. Με χρήση των βαρών λ_j που έχουν προκύψει, σταθμίζονται τα χαρακτηριστικά σε επίπεδο πλαισίου και στη συνέχεια υπολογίζονται τα σταθμισμένα στατιστικά για

κάθε ένα χαρακτηριστικό. Αποτέλεσμα είναι η χρήση μετασχηματισμένων χαρακτηριστικών ανά εκφώνηση, με σκοπό την ταξινόμηση με SVM.

3.5.3 Pre-MCE

Ως Pre-MCE, αναφέρεται η περίπτωση απλοϊκής ταξινόμησης κατά Bayes, με σκοπό την εκτίμηση των εκ-των-υστέρων πιθανοτήτων ανά ομάδα πλαισίων. Τα βάρη λ_j της Σχέσης 3.5.1 θεωρούνται ίσα με τη μονάδα. Με τον τρόπο αυτό, αγνοούνται τα βάρη που εκτιμώνται με την τεχνική MCE. Τελικά, επιλέγεται την κλάση με τη μεγαλύτερη εκ-των-υστέρων πιθανότητα δεδομένου μίας εκφώνησης. Η υλοποίηση του Απλοϊκού Bayes (Naive Bayes) ταξινομητή γίνεται με τη βοήθεια του εργαλείου MATLAB [41].

3.5.4 Post-MCE

Ως Post-MCE, αναφέρεται και το τελικό προτεινόμενο μοντέλο (*Late Fusion* στο Σχήμα 3.5). Οι εκ-των-υστέρων πιθανότητες ανά ομάδα πλαισίων προκύπτουν με υλοποίηση του Απλοϊκού Bayes ταξινομητή, και στη συνέχεια σταθμίζονται με χρήση των βαρών λ_j που έχουν προκύψει χάρη στο *Affective Saliency Model*. Οι πιθανότητες αυτές χρησιμοποιούνται για τον υπολογισμό της εκ-των-υστέρων πιθανότητας μιας εκφώνησης, σύμφωνα με τη Σχέση 3.5.1. Η τελική ταξινόμησή της γίνεται με βάση τη μεγαλύτερη εκ-των-υστέρων πιθανότητα, μεταξύ των διαφορετικών κλάσεων.

Κεφάλαιο 4

Κανονικοποίηση των Ακουστικών Χαρακτηριστικών

4.1 Εισαγωγή

Εφόσον τα χαρακτηριστικά που εξάγονται απευθείας από ανεπεξέργαστα δεδομένα μπορεί να έχουν μεγάλο εύρος τιμών και διαφορετικό μεταξύ τους, συχνά είναι απαραίτητη η κανονικοποίησή τους, πριν οποιαδήποτε χρήση τους από αλγορίθμους Μηχανικής Μάθησης. Για παράδειγμα, δεδομένου ότι ο υπολογισμός της Ευκλίδειας απόστασης μεταξύ δύο σημείων συναντάται συχνά σε τέτοιους αλγορίθμους, αξίζει να σημειωθεί ότι η τιμή της θα καθορίζεται σε σημαντικό βαθμό από ένα χαρακτηριστικό εάν αυτό έχει μεγαλύτερο εύρος τιμών από τα άλλα. Έτσι, μπορεί να θεωρηθεί κρίσιμο στοιχείο ενός συστήματος η κανονικοποίηση των χαρακτηριστικών (*Feature Normalization*), με σκοπό να είναι εύρωστο.

Όσον αφορά τομείς σχετικούς με την Επεξεργασία Φωνής, όπως η Αναγνώριση Φωνής ή Αναγνώριση Συναισθήματος από Φωνή, παρατηρείται μια επιπλέον μεταβολή στις τιμές των χαρακτηριστικών που εξάγονται από τα φωνητικά δεδομένα των διαφόρων ομιλητών. Η μεταβολή αυτή σχετίζεται με τη φύση του ήχου και οφείλεται σε μια σειρά παραγόντων, οι οποίοι κυρίως περιλαμβάνουν: αφενός τη διαφορετικότητα των ομιλητών και των συνθηκών ηχογράφησης, και αφετέρου τη συναισθηματική κατάσταση κάθε ομιλητή ή την πιθανή εξωτερίκευση κάποιου συναισθήματος. Ως διαφορετικότητα μεταξύ των ομιλητών αναφέρεται το σύνολο των στοιχείων εκείνων που διαφοροποιούν τα ακουστικά χαρακτηριστικά της φωνής τους. Ενδεικτικά, τέτοια στοιχεία μπορεί να είναι είτε βιολογικά, όπως η ηλικία, το φύλο και το ύψος, είτε κοινωνικοπολιτισμικά όπως η γλώσσα, η συμπεριφορά και ο προσωπικός χαρακτήρας, όπως αναλύθηκε και στο Κεφάλαιο 2. Αντίστοιχα, οι συνθήκες ηχογράφησης μπορεί να επηρεαστούν από την ακουστική του χώρου, το διαφορετικό είδος μικροφώνου ή θέση του ομιλητή, αλλά και από το signal-to-noise ratio (SNR). Όλα τα παραπάνω σε συνδυασμό με τη συναισθηματική έκφραση, επηρεάζουν σε σημαντικό βαθμό τα εξαγόμενα χαρακτηριστικά και κατ' επέκταση την απόδοση του συστήματος αναγνώρισης, είτε αυτά προορίζονται για εκπαίδευση είτε για αξιολόγηση. Επίσης, αναντιστοιχία είναι πιθανόν να παρατηρηθεί μεταξύ των δεδομένων εκπαίδευσης και αξιολόγησης, όπως σε πραγματικού χρόνου εφαρμογές.

Μια πρώτη προσέγγιση είναι, λοιπόν, η κανονικοποίηση των ακουστικών χαρακτηριστικών, η οποία μπορεί να υλοποιηθεί με μια σειρά από τεχνικές. Οι κυριότερες αναφέρονται στις παρα-

γράφους που ακολουθούν. Ονομαστικά, κάποιες από αυτές είναι οι: *Z-Normalization*, *Peak-to-Peak Normalization* και *Percentile Normalization*. Πρέπει να σημειωθεί ότι κάθε τεχνική κανονικοποίησης μπορεί να εφαρμοστεί αναφορικά με τις τιμές ενός συγκεκριμένου χαρακτηριστικού: σε όλα τα πλαίσια όλων των διαθέσιμων εκφωνήσεων (καθολική κανονικοποίηση ή *global normalization*), στα πλαίσια των εκφωνήσεων κάθε ομιλητή ξεχωριστά (εξαρτώμενη-του-ομιλητή ή *speaker-dependent*), στα πλαίσια των συνολικών εκφωνήσεων κάθε βάσης δεδομένων (εξαρτώμενη-της-βάσης ή *corpus-dependent*) ή στα πλαίσια κάθε εκφώνησης (εξαρτώμενη-της-εκφώνησης ή *utterance-based*). Συνήθως, η καθολική κανονικοποίηση έχει χειρότερη απόδοση από τις άλλες τρεις, καθώς στον υπολογισμό της μπορεί να περιλαμβάνει δεδομένα από διαφορετικούς ομιλητές, περιβάλλοντα ή βάσεις δεδομένων. Στην περίπτωση όπου οι διαθέσιμες εκφωνήσεις προέρχονται από μόνο μία βάση, τότε η εξαρτώμενη-της-βάσης κανονικοποίηση ισοδυναμεί με την καθολική.

4.2 Τεχνικές Κανονικοποίησης των Χαρακτηριστικών

4.2.1 Peak Normalization

Μια απλή τεχνική κανονικοποίησης των χαρακτηριστικών είναι η Peak Normalization [42]. Εφαρμόζεται με αφαίρεση της μέγιστης τιμής ενός χαρακτηριστικού από την τιμή του σε κάθε πλαίσιο. Έστω x_k η τιμή του χαρακτηριστικού x στο k-οστό πλαίσιο, τότε η νέα τιμή $\hat{x}_k$ προκύπτει:

$$\hat{x}_k = x_k - x_{max} \tag{4.2.1}$$

Το αποτέλεσμα είναι η κλιμάκωση των χαρακτηριστικών, έτσι ώστε να έχουν μέγιστη τιμή ίση με 0. Όπως αναφέρθηκε στην εισαγωγή, ο παραπάνω υπολογισμός μπορεί να εφαρμοστεί για τα πλαίσια κάθε εκφώνησης ή κάθε ομιλητή ξεχωριστά, ή ακόμα και για όλα τα διαθέσιμα πλαίσια. Ανάλογα, προκύπτει και η μέγιστη τιμή $x_{max} = \max_k\{x_k\}$.

4.2.2 Peak-to-Peak Normalization

Αντίστοιχα με την προηγούμενη τεχνική, είναι δυνατή η απεικόνιση των χαρακτηριστικών μεταξύ μιας μέγιστης και μιας ελάχιστης τιμής. Συνήθως, ορίζεται το διάστημα [0,1] ή [-1,1]. Η συγκεκριμένη τεχνική ονομάζεται Peak-to-Peak Normalization, αλλά συναντάται και με το όνομα Min-Max Normalization [43, 44, 45]. Θεωρώντας το εύρος τιμών ενός χαρακτηριστικού x ίσο με το διάστημα $[x_{min},x_{max}]$ και το ζητούμενο διάστημα απεικόνισης $[y_{min},y_{max}]$, τότε η κανονικοποιημένη τιμή του χαρακτηριστικού στο k-οστό πλαίσιο υπολογίζεται σύμφωνα με τη σχέση:

$$\hat{x}_k = (y_{max} - y_{min}) * \frac{x_k - x_{min}}{x_{max} - x_{min}} + y_{min} \tag{4.2.2}$$

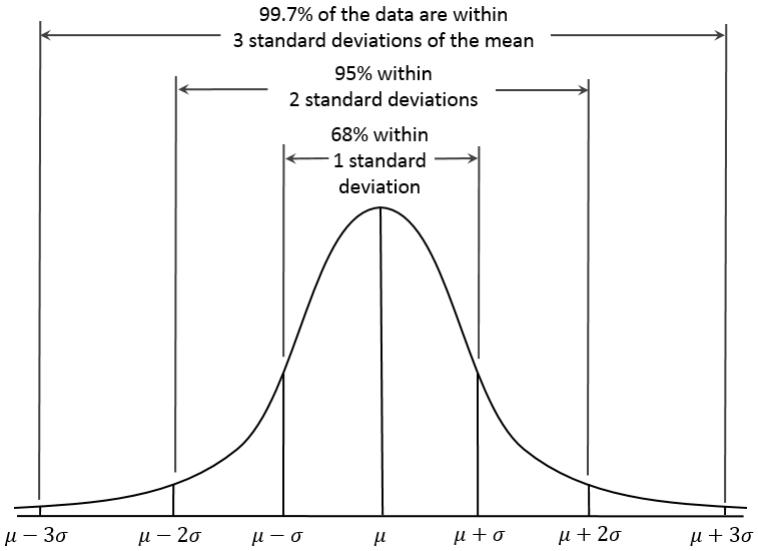
Για τον υπολογισμό των x_{min}, x_{max}, όπως αναφέρθηκε και παραπάνω, μπορούν να περιληφθούν τα πλαίσια μιας εκφώνησης, ενός ομιλητή, μίας βάσης δεδομένων ή και τα συνολικά.

4.2.3 Z-Normalization

Η Z-Normalization είναι μια συνήθης τεχνική [44, 46, 47, 48, 49, 50], με σκοπό τα δείγματα να αποκτήσουν μέση τιμή 0 και τυπική απόκλιση 1. Μετά την εφαρμογή της, διατηρούνται ιδιότητες σχετικά με το σχήμα της αρχικής κατανομής των δεδομένων, όπως η κύρτωση (kurtosis) και η λοξότητα (skewness) [41]. Έστω μ η μέση τιμή των δειγμάτων ενός χαρακτηριστικού x κατά μήκος των πλαισίων μιας εκφώνησης ή των εκφωνήσεων ενός ομιλητή, ή ακόμα και των συνολικών εκφωνήσεων (ανάλογα με τον τρόπο που ακολουθείται), και σ η αντίστοιχη τυπική απόκλιση, τότε η κανονικοποίηση γίνεται με εφαρμογή της σχέσης:

$$\hat{x}_k = \frac{x_k - \mu}{\sigma} \tag{4.2.3}$$

Η τιμή που προκύπτει, αντιπροσωπεύει ουσιαστικά την απόσταση της αρχικής τιμής x_k από την αρχική μέση τιμή, μετρημένη σε μονάδες τυπικής απόκλισης. Αν το $\hat{x}_k$ είναι αρνητικό, τότε το x_k είναι μικρότερο της μέσης τιμής μ, ενώ αν το $\hat{x}_k$ είναι θετικό, τότε το x_k είναι μεγαλύτερο. Αξίζει να σημειωθεί ότι εάν τα δεδομένα ακολουθούν κανονική κατανομή, μόνο το 5% των δειγμάτων θα βρίσκεται εκτός του διαστήματος $[-2\sigma,2\sigma]$, σύμφωνα και με το [32].



Σχήμα 4.1: *Κανονική κατανομή (Normal Distribution). (Η γραφική απεικόνιση προέρχεται από το [51].)*

4.2.4 Mean Normalization

Μια ακόμα συνήθης τακτική είναι η αφαίρεση της μέσης τιμής μ των δειγμάτων ενός χαρακτηριστικού, υπολογισμένης αντίστοιχα με παραπάνω. Σε αυτήν την περίπτωση, εφαρμόζεται η σχέση:

$$\hat{x}_k = x_k - \mu \tag{4.2.4}$$

Τα κανονικοποιημένα χαρακτηριστικά θα αποκτήσουν μέση τιμή 0 (*Zero-Mean Normalization* [52]). Σκοπός είναι η εξάλειψη οποιουδήποτε επιπλέον παράγοντα, ο οποίος συμβάλλει σε μεταβολή των τιμών του χαρακτηριστικού, όπως οι διαφορετικές συνθήκες ηχογράφησης.

Η συγκεκριμένη τεχνική συναντάται τις περισσότερες φορές ως *Cepstral Mean Normalization* [42, 53, 54], όπου εφαρμόζεται αφαίρεση της μέσης τιμής σε φασματικά χαρακτηριστικά, όπως τα MFCC. Η περίπτωση αυτή αντιμετωπίζεται και ως τεχνική προσαρμογής ενός συστήματος σε καινούργιο ακουστικό περιβάλλον. Για το λόγο αυτό, αναλύεται στο κεφάλαιο 5.2.

Επίσης, είναι δυνατή η αφαίρεση της μέσης τιμής από χαρακτηριστικό υπολογισμένο σε λογαριθμική κλίμακα, όπως η log-Ενέργεια [55]. Η περίπτωση αυτή αντιστοιχεί, ουσιαστικά, σε διαίρεση του χαρακτηριστικού (της Ενέργειας) με τη μέση τιμή του. Εδώ, όμως, θα θεωρηθεί ως είδος της τεχνικής *Mean Normalization*.

4.2.5 Percentile Normalization

Αντίστοιχα με την τεχνική αφαίρεσης της μέγιστης τιμής ενός χαρακτηριστικού (Peak Normalization), είναι δυνατή η αφαίρεση του p-οστού εκατοστημορίου (percentile). Το p-οστό εκατοστημόριο των δειγμάτων ενός χαρακτηριστικού αντιστοιχεί στην τιμή εκείνη, η οποία είναι μεγαλύτερη από το p% των δειγμάτων. Όσον αφορά την τιμή του p, μια πιθανή τιμή είναι το 90, με την έννοια ότι πάνω από το 90-οστό εκατοστημόριο υπάρχουν δείγματα που αποτελούν εξαιρέσεις της συνολικής κατανομής του συγκεκριμένου χαρακτηριστικού (outliers). Σύμφωνα με τα παραπάνω, η συγκεκριμένη τεχνική κανονικοποίησης μπορεί να εφαρμοστεί σύμφωνα με τη σχέση:

$$\hat{x}_k = x_k - n_{90} \tag{4.2.5}$$

όπου ως n_{90} έχει οριστεί η τιμή, η οποία είναι μεγαλύτερη από το 90% των δειγμάτων του χαρακτηριστικού. Το αποτέλεσμα είναι το 90-οστό εκατοστημόριο των κανονικοποιημένων χαρακτηριστικών να παίρνει την τιμή 0.

4.2.6 Percentile Peak-to-Peak Normalization

Αντίστοιχα με την τεχνική Peak-to-Peak Normalization, η οποία απεικονίζει τις τιμές των δειγμάτων ενός χαρακτηριστικού μεταξύ ενός καθορισμένου διαστήματος, είναι δυνατή η απεικόνιση μέρους των δειγμάτων αυτών στο συγκεκριμένο διάστημα. Με δεδομένο ότι οι τιμές ενός χαρακτηριστικού πάνω από το 90-οστό εκατοστημόριο και κάτω από το 10-ο εκατοστημόριο, αποτελούν εξαιρέσεις, μπορεί να κανονικοποιηθεί το σύνολο των τιμών, έτσι ώστε το 90-οστό εκατοστημόριο (n_{90}) να πάρει την τιμή y_{max} και το 10-οστό εκατοστημόριο (n_{10}) να πάρει την τιμή y_{min}. Η κανονικοποίηση αυτή εφαρμόζεται σύμφωνα με τη σχέση:

$$\hat{x}_k = (y_{max} - y_{min}) * \frac{x_k - n_{10}}{n_{90} - n_{10}} + y_{min} \tag{4.2.6}$$

Ως διάστημα τιμών [y_{min},y_{max}] συνήθως θεωρείται το [0,1] ή το [-1,1].

4.2.7 Histogram Equalization

Η τεχνική Histogram Equalization έχει αναπτυχθεί στον τομέα της Επεξεργασίας Εικόνας [56]. Αποσκοπεί στη βελτίωση της αντίθεσης και της φωτεινότητας μιας εικόνας, μέσω της αλλαγής του εύρους τιμών των pixels της. Ο μετασχηματισμός που επιβάλλει είναι μη γραμμικός, θεωρώντας την αθροιστική συνάρτηση πυκνότητας πιθανότητας της αρχικής εικόνας. Ουσιαστικά,

η τεχνική αυτή μετασχηματίζει το ιστόγραμμα της αρχικής εικόνας, έτσι ώστε να προσεγγίσει ένα ιστόγραμμα αναφοράς. Ως αναφορά, συνήθως, χρησιμοποιείται η ομοιόμορφη κατανομή.

Δοκιμές έχουν γίνει για την εφαρμογή της συγκεκριμένης τεχνικής και στην Αναγνώριση Φωνής [57, 58]. Στην περίπτωση αυτή, δείχνει να μειώνει σημαντικά τη μη γραμμική παραμόρφωση ενός σήματος εξαιτίας αυξημένου θορύβου. Έτσι, συνεισφέρει σε μείωση της αναντιστοιχίας των δεδομένων εκπαίδευσης και αναγνώρισης, αν αυτά προέρχονται από διαφορετικές συνθήκες ηχογράφησης.

4.3 Εφαρμογή των Τεχνικών Κανονικοποίησης

Στο παρόν κεφάλαιο πραγματοποιείται εφαρμογή και σύγκριση των παραπάνω τεχνικών κανονικοποίησης των χαρακτηριστικών. Για το σκοπό αυτό, υλοποιούνται τα 4 μοντέλα (*Baseline*, *Early Fusion*, *Pre-MCE*, *Post-MCE*) που περιγράφηκαν στην παράγραφο 3.5, με χρήση του *Affective Saliency Model*. Χρησιμοποιείται ο έτοιμος κώδικας των μοντέλων [31], με επιπλέον υλοποίηση κάθε τεχνικής κανονικοποίησης ξεχωριστά. Πριν την ανάλυση των αποτελεσμάτων, παρατίθενται λεπτομέρειες σχετικά με τα δεδομένα και τα ακουστικά χαρακτηριστικά που εξάχθηκαν.

4.3.1 Δεδομένα

Η βάση δεδομένων που χρησιμοποιείται στο παρόν κεφάλαιο είναι η LEGO (Let's Go Bus Information System από το Carnegie Mellon University), και συγκεκριμένα ένα μέρος των δεδομένων της [59]. Τα συγκεκριμένα δεδομένα αποτελούνται από πληροφορίες που παρέχονται στους πολίτες, όσον αφορά δρομολόγια λεωφορείων στην πόλη Pittsburgh, από ένα κατάλληλα διαμορφωμένο διαλογικό σύστημα. Οι ετικέτες, που έχουν δοθεί στις εκφωνήσεις και αφορούν επίπεδα θυμού, ακολουθούν τα 5 παρακάτω επίπεδα: φιλική, ουδέτερη, λίγο θυμωμένη, θυμωμένη, πολύ θυμωμένη. Κάθε εκφώνηση έχει διάρκεια περίπου 1-2 δευτερόλεπτα. Παρόμοια με το [31], χρησιμοποιούνται 4243 εκφωνήσεις, οι οποίες χωρίζονται στις δύο παρακάτω κλάσεις:

- *Μη-θυμωμένη* (3309 εκφωνήσεις): φιλική, ουδέτερη
- *Θυμωμένη* (934 εκφωνήσεις): λίγο θυμωμένη, θυμωμένη, πολύ θυμωμένη

4.3.2 Ακουστικά Χαρακτηριστικά

Ως ακουστικά χαρακτηριστικά, με σκοπό την εφαρμογή και σύγκριση των παραπάνω τεχνικών κανονικοποίησης, θεωρείται το σύνολο των χαρακτηριστικών που αναφέρεται στο [31]. Η εξαγωγή τους έγινε με τη βοήθεια του εργαλείου Opensmile [60]. Κάθε εκφώνηση χωρίζεται σε πλαίσια διάρκειας 30 msec, με μετακίνηση του πλαισίου κατά 10 msec κάθε φορά. Συνολικά, εξάγονται 33 ακουστικά χαρακτηριστικά για κάθε πλαίσιο, για τα οποία επιπλέον υπολογίζεται η πρώτη παράγωγός τους. Παρακάτω αναγράφονται τα χαρακτηριστικά αυτά, ανάλογα με την κατηγορία που ανήκουν:

- *Σχετικά με την Ενέργεια*: Ενέργεια, Ρυθμός διέλευσης από το μηδέν

- *Φασματικά*: Ενέργεια 250-650 Hz και 1-4 kHz, Φασματική Ροή, Εντροπία, Διακύμανση, Κύρτωση, Λοξότητα, Κλίση, Ψυχοακουστική Οξύτητα, Αρμονικότητα, MFCC συνιστώσες 1-14, Συχνότητα κάτω από την οποία βρίσκεται το 25%, 50%, 75%, 90% των συχνοτήτων (Roll-off Point 0.25, 0.50, 0.75, 0.90)

- *Φωνητικά*: Θεμελιώδης Συχνότητα (Pitch), Ανεπεξέργαστη Θεμελιώδης Συχνότητα (Raw Pitch), Πιθανότητα Φωνής

Για τον υπολογισμό των χαρακτηριστικών σε επίπεδο εκφώνησης, επιβάλλονται στα παραπάνω χαρακτηριστικά τα 5 στατιστικά: ελάχιστη τιμή, μέγιστη τιμή, μέση τιμή, διάμεσος και τυπική απόκλιση, όπως και στο [31]. Με τον τρόπο αυτό, προκύπτουν 330 χαρακτηριστικά ανά εκφώνηση, τα οποία χρησιμοποιούνται για την εκπαίδευση ή αξιολόγηση SVM ταξινομητή, όπου αυτός υλοποιείται.

Όσον αφορά τον Απλοϊκό Bayes ταξινομητή, από τα παραπάνω χαρακτηριστικά ανά πλαίσιο επιλέγονται τα εξής τρία: ενέργεια, πρώτη MFCC συνιστώσα, ανεπεξέργαστη θεμελιώδης συχνότητα. Στη συνέχεια, ομαδοποιούνται σε ομάδες των 20 πλαισίων (segments), ξεχωριστά για κάθε εκφώνηση. Για κάθε τέτοια ομάδα, υπολογίζονται τα ίδια 5 στατιστικά με πριν (ελάχιστη τιμή, μέγιστη τιμή, μέση τιμή, διάμεσος και τυπική απόκλιση). Με τον τρόπο αυτό προκύπτουν 15 χαρακτηριστικά ανά ομάδα, με σκοπό την εκπαίδευση ή αξιολόγηση Απλοϊκού Bayes ταξινομητή.

Επιπρόσθετα, εξάγονται η φασματική ροή και η θεμελιώδης συχνότητα με σταθερό πλαίσιο μήκους 200 msec. Τα δύο αυτά χαρακτηριστικά, σε συνδυασμό με τη θεμελιώδη συχνότητα που έχει εξαχθεί με πλαίσιο μήκους 30 msec και μετακίνηση κατά 10 msec, χρησιμοποιούνται ως *Regression Features*. Επιπλέον, υπολογίζεται ο ρυθμός άφωνων πλαισίων ανά ομάδα, με χρήση Voice Activity Detector [39]. Για τη θεμελιώδη συχνότητα που έχει εξαχθεί με πλαίσιο μήκους 30 msec και μετακίνηση κατά 10 msec, τα πλαίσια ομαδοποιούνται σε ομάδες των 20 πλαισίων, όπως και πριν, και ως χαρακτηριστικά καταχωρούνται τα 5 στατιστικά.

4.3.3 Αποτελέσματα

Με χρήση των 4 μοντέλων (*Baseline*, *Early Fusion*, *Pre-MCE*, *Post-MCE*), εφαρμόστηκε τόσο καθολική κανονικοποίηση των χαρακτηριστικών, όσο και κανονικοποίηση ανά εκφώνηση, με κάθε μέθοδο ξεχωριστά. Γενικά, η καθολική κανονικοποίηση δεν είναι ιδιαίτερα αποτελεσματική, όπως αναφέρθηκε και στην εισαγωγή. Αυτό οφείλεται στη χρήση του συνόλου των δεδομένων για την εκτίμηση των παραμέτρων κανονικοποίησης, όπου τα δεδομένα είναι δυνατόν να προέρχονται από διαφορετικούς ομιλητές και ακουστικά περιβάλλοντα. Έτσι, αποτελεσματικότερη κρίνεται η κανονικοποίηση ανά ομιλητή (βλ. κεφ. 6) ή η κανονικοποίηση ανά εκφώνηση, έτσι ώστε να μειωθούν οι διαφορές αυτές. Στο παρόν κεφάλαιο εξετάζεται η δεύτερη περίπτωση. Τελικά, όπως αναμενόταν, συγκρίνοντας με την καθολική κανονικοποίηση, τα αποτελέσματα της κανονικοποίησης ανά εκφώνηση πρόεκυψαν καλύτερα.

Ενδεικτικά, παρακάτω παρατίθενται τα ποσοστά επιτυχίας σε περίπτωση καθόλου κανονικοποίησης, καθολικής κανονικοποίησης με τη μέθοδο Peak-to-Peak Normalization και κανονικοποίησης ανά εκφώνηση με την ίδια μέθοδο. Τα ποσοστά αυτά αφορούν το μοντέλο Post-MCE, το οποίο ήταν και το προτεινόμενο του αντίστοιχου άρθρου [31]. Παρατηρείται μικρή βελτίωση, σύμφωνα με την παραπάνω ανάλυση. Γενικά, λόγω των δεδομένων της συγκεκριμένης

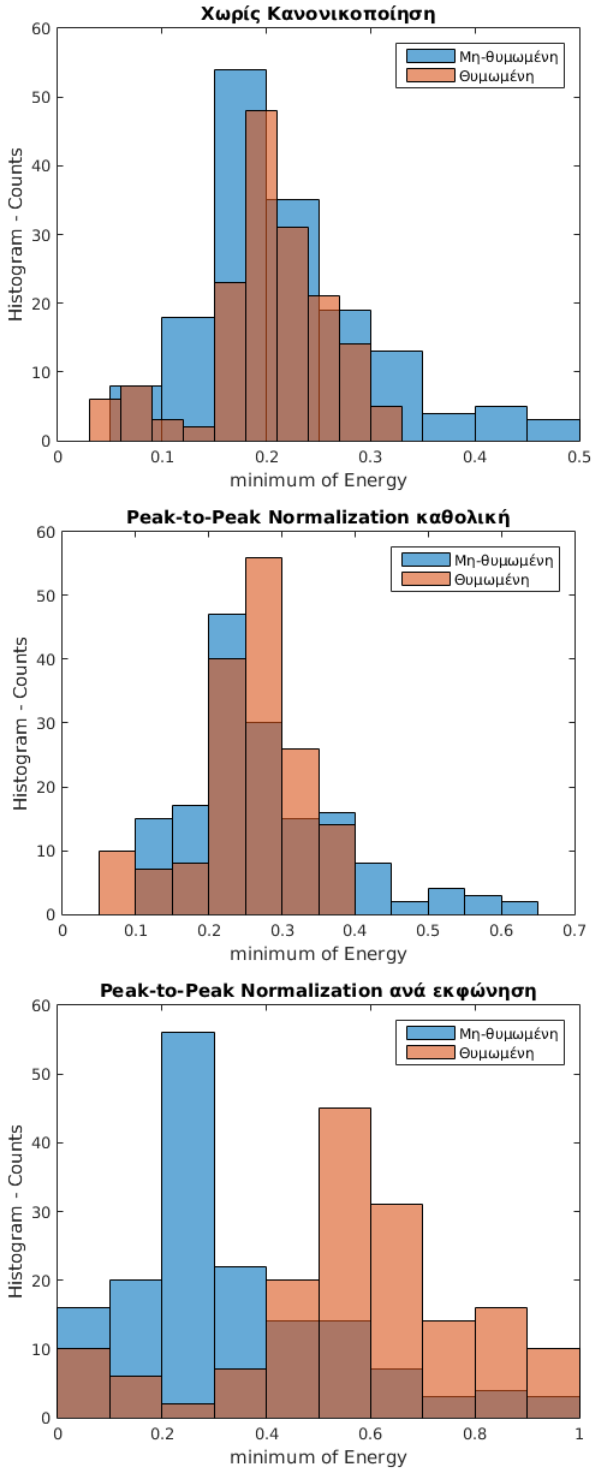
βάσης, τα οποία προέρχονται από ηχογράφηση του ίδιου διαλογικού συστήματος σε λειτουργία, δεν παρατηρούνται σημαντικά διαφορετικές συνθήκες ηχογράφησης. Αποτέλεσμα είναι να μην βελτιώνεται σημαντικά η απόδοση των μοντέλων, κατά την εφαρμογή κανονικοποίησης ανά εκφώνηση.

	Post-MCE
Χωρίς κανονικοποίηση	79.4
Καθολική Peak-to-Peak ([0,1])	79.4
Ανά Εκφώνηση Peak-to-Peak ([0,1])	**79.9**

Πίνακας 4.1: *Ποσοστά επιτυχίας (%) για τις 3 περιπτώσεις: Χωρίς κανονικοποίηση, Καθολική κανονικοποίηση και Κανονικοποίηση ανά εκφώνηση. Ως τεχνική κανονικοποίησης χρησιμοποιείται η Peak-to-Peak Normalization στο διάστημα [0,1].*

Επιπρόσθετα, παρατηρήθηκε ότι σε κάθε μοντέλο λειτουργεί καλύτερα διαφορετική τεχνική. Για παράδειγμα, όσον αφορά την καθολική κανονικοποίηση, στο Baseline το μεγαλύτερο ποσοστό επιτυχίας έδωσε η Peak-to-Peak στο [-1,1] (80.2%), ενώ στο Early Fusion έδωσε η Z-Normalization (80.6%).

Στο Σχήμα 4.2, απεικονίζονται τα ιστογράμματα του χαρακτηριστικού Ελάχιστη τιμή της Ενέργειας, για δύο εκφωνήσεις που ανήκουν αντίστοιχα στη Μη-θυμωμένη (μπλε) και στη Θυμωμένη (κόκκινη) κλάση. Παρατίθενται και οι τρεις περιπτώσεις κανονικοποίησης: Χωρίς κανονικοποίηση, Καθολική κανονικοποίηση και Κανονικοποίηση ανά εκφώνηση. Ως τεχνική κανονικοποίησης χρησιμοποιείται η Peak-to-Peak Normalization στο διάστημα [0,1], αντίστοιχα με τα προηγούμενα ποσοστά επιτυχίας. Παρατηρείται σημαντική διαφορά μεταξύ της καθολικής κανονικοποίησης και της κανονικοποίησης ανά εκφώνηση, καθώς στην πρώτη περίπτωση οι κατανομές επικαλύπτονται σε μεγάλο βαθμό. Στην κανονικοποίηση ανά εκφώνηση, αντίθετα, έχει επιτευχθεί καλύτερος διαχωρισμός των δύο κλάσεων.



Σχήμα 4.2: *Ιστογράμματα του χαρακτηριστικού Ελάχιστη τιμή της Ενέργειας, για 2 εκφωνήσεις (Μη-θυμωμένη και Θυμωμένη κλάση), για τις 3 περιπτώσεις: Χωρίς κανονικοποίηση, Καθολική κανονικοποίηση και Κανονικοποίηση ανά εκφώνηση. Ως τεχνική κανονικοποίησης χρησιμοποιείται η Peak-to-Peak Normalization στο διάστημα [0,1].*

Κεφάλαιο 5

Τεχνικές Προσαρμογής του Ομιλητή

5.1 Εισαγωγή

Για την επίτευξη όσο το δυνατόν μικρότερου ποσοστού λάθους, είναι απαραίτητη η προσαρμογή ενός συστήματος αναγνώρισης φωνής σε έναν καινούργιο ομιλητή. Όπως αναφέρθηκε και στο Κεφάλαιο 1, κάθε ομιλητής χαρακτηρίζεται από κάποια βιολογικά και κοινωνικο-πολιτισμικά στοιχεία, τα οποία διαφοροποιούν σημαντικά τη φωνή του, σε σχέση με τους υπόλοιπους ομιλητές. Επιπλέον, ιδιαίτερη επιρροή έχουν και οι συνθήκες ηχογράφησης. Εκτός, λοιπόν, από την κανονικοποίηση των εξαγόμενων χαρακτηριστικών του κάθε ομιλητή, ένα δεύτερο βήμα για τη βελτίωση ενός συστήματος είναι η προσαρμογή του σε αυτό. Η προσαρμογή αυτή θα μπορούσε να γίνει με εκπαίδευση του συστήματος, κάθε φορά, με δεδομένα αποκλειστικά από το συγκεκριμένο ομιλητή (*speaker-dependent* μοντέλο). Αυτό, όμως, θα απαιτούσε ένα πολύ μεγάλο όγκο δεδομένων και μια πολύ χρονοβόρα διαδικασία. Αντίθετα, μπορούν να εκπαιδευτούν συστήματα ανεξάρτητα του ομιλητή (*speaker-independent*), χρησιμοποιώντας δεδομένα από μια ποικιλία διαφορετικών ομιλητών. Αυτά τα μοντέλα δίνουν καλύτερα αποτελέσματα σε περίπτωση καινούργιου ομιλητή, αλλά συνολικά έχουν χαμηλότερη απόδοση. Μια προσέγγιση είναι η κατάλληλη προσαρμογή τους κάθε φορά στον ομιλητή (*speaker-adaptive* μοντέλα), χρησιμοποιώντας δεδομένα από λίγα σχετικά δείγματα φωνής. Στον τομέα της Αυτόματης Αναγνώρισης Φωνής (ASR), έχουν αναπτυχθεί μία σειρά από τεχνικές για αυτή την προσαρμογή (*Speaker Adaptation*). Οι κυριότερες από αυτές αναλύονται παρακάτω.

Ιδιαίτερο ενδιαφέρον παρουσιάζει το γεγονός ότι έχει γίνει προσπάθεια για την εφαρμογή των παρακάτω τεχνικών για την προσαρμογή ομιλητή και σε άλλους τομείς εκτός από την αναγνώριση φωνής. Τέτοιοι τομείς αφορούν, παραδείγματος χάριν, την αναγνώριση ομιλητή, ηλικίας αλλά και συναισθήματος από φωνή. Ανεξάρτητα του τομέα, στόχος είναι η προσαρμογή του συστήματος στα δεδομένα αξιολόγησης, έτσι ώστε να επιτευχθεί μεγαλύτερο ποσοστό επιτυχίας.

Γενικά, οι τεχνικές προσαρμογής ενός ομιλητή μπορούν να εφαρμοστούν με επίβλεψη (*supervised*) ή χωρίς επίβλεψη (*unsupervised*). Μια μέθοδος θεωρείται ότι εφαρμόζεται με επίβλεψη όταν οι ετικέτες (ή οι μεταγραφές, ανάλογα με τον τομέα που αναφέρεται το σύστημα) των δεδομένων εκπαίδευσης είναι γνωστές. Αντίθετα, χωρίς επίβλεψη θεωρείται όταν οι ετικέτες είναι άγνωστες. Προφανώς, ένα χωρίς επίβλεψη εκπαιδευμένο μοντέλο έχει χαμηλότερη απόδοση. Όμως, έχει μεγαλύτερο ενδιαφέρον καθώς αντιστοιχεί καλύτερα σε πραγματικού χρόνου

καταστάσεις, όπου δεν υπάρχει ο χρόνος για την επισημείωση των δεδομένων.

Ένας ακόμα διαχωρισμός που μπορεί να γίνει αφορά τη φύση της προσαρμογής: αν είναι στατική (*static*) ή δυναμική (*dynamic*). Στατική ονομάζεται όταν δίνονται στο σύστημα όλα τα δεδομένα με σκοπό την προσαρμογή του. Αντίθετα, δυναμική μπορεί να θεωρηθεί όταν τα δεδομένα είναι σταδιακά διαθέσιμα, όπως στην περίπτωση ενός διαλογικού συστήματος (*spoken dialogue system*) ή άλλης πραγματικού χρόνου εφαρμογής.

Τέλος, προσαρμογή του ομιλητή μπορεί να επιτευχθεί με μετασχηματισμό των δεδομένων/χαρακτηριστικών του (*feature-based*) ή με μετασχηματισμό των παραμέτρων του ανεξάρτητου-του-ομιλητή μοντέλου (*model-based*). Άλλες τεχνικές κατατάσσονται στη μία ή στην άλλη κατηγορία, ενώ κάποιες μπορούν να εφαρμοστούν και με τους δύο τρόπους. Ιδιαίτερο ενδιαφέρον έχει η τεχνική προσαρμοζόμενης εκπαίδευσης, επονομαζόμενη ως *Speaker Adaptive Training (SAT)*, όπου με χρήση κάποιας τεχνικής, στοχεύει ήδη από την εκπαίδευση στην καλύτερη προσαρμογή του μοντέλου.

Στις παραγράφους που ακολουθούν αναλύονται οι κυριότερες τεχνικές για προσαρμογή του ομιλητή που χρησιμοποιούνται στον τομέα της Αναγνώρισης Φωνής: *Cepstral Mean Normalization (CMN)*, *Vocal Tract Length Normalization (VTLN)*, *Maximum A-Posteriori (MAP) Adaptation*, *Maximum Likelihood Linear Regression (MLLR)*, *Speaker Adaptive Training (SAT)* (παρ. 5.2, 5.3, 5.4, 5.5, 5.6 αντίστοιχα). Συμπληρωματικά, στις επιπλέον υποπαραγράφους, αναφέρονται κάποιες λεπτομέρειες όσον αφορά την υλοποίηση της κάθε τεχνικής που έχει αναπτύξει το Kaldi [61] (παρ. 5.3.1, 5.3.2, 5.5.1). Το Kaldi αποτελεί ένα εργαλείο αναγνώρισης φωνής, για το οποίο μια συνοπτική περιγραφή παρέχεται στο Παράρτημα Α.

5.2 Cepstral Mean Normalization (CMN)

Η Cepstral Mean Normalization (CMN) εφαρμόζεται με σκοπό τη μείωση της διαφοράς μεταξύ διαφορετικών ακουστικών περιβαλλόντων. Θεωρώντας ότι το ακουστικό περιβάλλον είναι σταθερό συγκριτικά με το σήμα φωνής, η κρουστική του απόκριση μπορεί να εκτιμηθεί ως η μέση τιμή του cepstrum μιας σειράς πλαισίων. Έτσι, είναι δυνατή η αφαίρεσή του με εφαρμογή της σχέσης:

$$\hat{c}(n,k) = c(n,k) - \overline{c}(n) \tag{5.2.1}$$

όπου k ο αριθμός του πλαισίου και $c(n) = |\mathcal{F}\{log|\mathcal{F}\{s(n)\}|^2\}|^2$ το power cepstrum του σήματος $s(n)$, που ορίζεται ως το power spectrum του λογαρίθμου του power spectrum [54].

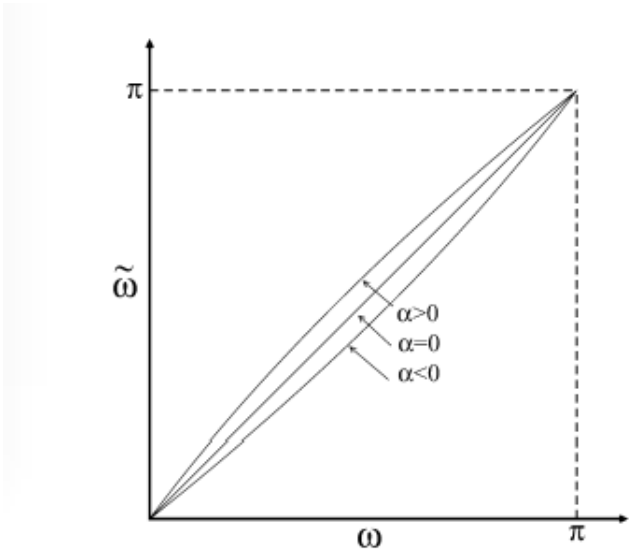
Στην πράξη, η CMN εφαρμόζεται σε φασματικά χαρακτηριστικά, όπως τα MFCCs. Για κάθε MFCC χαρακτηριστικό, υπολογίζεται η μέση τιμή του σε όλα τα πλαίσια μιας εκφώνησης ή ενός ομιλητή. Η μέση τιμή αυτή αφαιρείται από την τιμή του συγκεκριμένου χαρακτηριστικού σε κάθε πλαίσιο της εκφώνησης ή του ομιλητή αντίστοιχα. Με αυτό τον τρόπο, το χαρακτηριστικό κανονικοποιείται, έτσι ώστε να έχει μέση τιμή 0. Επιπλέον κανονικοποίηση, μπορεί να γίνει και με διαίρεση του χαρακτηριστικού κάθε πλαισίου με την αντίστοιχη τυπική απόκλιση. Τότε, αναφέρεται ως *Cepstral Mean and Variance Normalization* (CMVN). Τα κανονικοποιημένα χαρακτηριστικά θα έχουν μέση τιμή 0 και διακύμανση 1.

5.3 Vocal Tract Length Normalization (VTLN)

Μια συχνά χρησιμοποιούμενη τεχνική στον τομέα της Αναγνώρισης Φωνής (ASR), με σκοπό την ελαχιστοποίηση των διαφορών μεταξύ των ομιλητών (*inter-speaker variability*), είναι η Vocal Tract Length Normalization (VTLN). Κάθε ομιλητής έχει διαφορετικό μήκος φωνητικής οδού (Vocal Tract Length ή VTL), που εξαρτάται από την ηλικία, το φύλο και το μεγέθος του σώματός του. Το μήκος αυτό επηρεάζει σημαντικά το φάσμα του σήματος φωνής. Ιδιαίτερη μεταβολή παρατηρείται στις θέσεις των συχνοτήτων φωνοσυντονισμού, οι οποίες είναι αντιστρόφως ανάλογες του συγκεκριμένου μήκους. Ενδεικτικά, οι συχνότητες φωνοσυντονισμού είναι κατά 20% υψηλότερες στις γυναίκες από τους άντρες [62]. Η βασική ιδέα της VTLN είναι ο μετασχηματισμός του άξονα των συχνοτήτων f κάθε ομιλητή, έτσι ώστε να μειωθούν οι φασματικές διαφορές μεταξύ τους. Διάφορες συναρτήσεις στρέβλωσης (*warping functions*) g_α έχουν χρησιμοποιηθεί με σκοπό αυτό το μετασχηματισμό:

$$\hat{f} = g_\alpha(f) \tag{5.3.1}$$

όπου το α αποτελεί σταθερά για κάθε συνάρτηση g_α, ονομάζεται παράγοντας στρέβλωσης (*warping factor*) και προσδιορίζεται για κάθε έναν από τους ομιλητές. Παρακάτω απεικονίζεται ένα παράδειγμα γραφικής παράστασης της κανονικοποιημένης συχνότητας $\hat{\omega} = g_\alpha(\omega)$ σε rad (Σχήμα 5.1).



Σχήμα 5.1: *Παράδειγμα μιας VTLN συνάρτησης στρέβλωσης (warping function)* $\hat{\omega} = g_\alpha(\omega)$ *για διαφορετικές τιμές του* α*. (Η γραφική απεικόνιση προέρχεται από το [62].)*

Υπάρχουν 2 προσεγγίσεις της VTLN για την εκτίμηση του παράγοντα α: με βάση το ηχητικό σήμα (*signal-based*) και με βάση το μοντέλο (*model-based*). Στην πρώτη προσέγγιση απαιτείται, συνήθως, ακριβής υπολογισμός των συχνοτήτων φωνοσυντονισμού. Ακολουθεί η εκτίμηση του παράγοντα α, η οποία μπορεί να γίνει με το παρακάτω σύστημα εξισώσεων, σύμφωνα με το [54]:

$$\begin{aligned} F_{1,ref} &= \alpha F_1 \\ F_{2,ref} &= \alpha F_2 \end{aligned} \tag{5.3.2}$$

όπου F_1, F_2 η πρώτη και δεύτερη συχνότητα φωνοσυντονισμού του νέου ομιλητή και $F_{1,ref}$, $F_{2,ref}$ οι αντίστοιχες συχνότητες ενός εκπαιδευμένου μοντέλου αναφοράς.

Η model-based VTLN εκπαιδεύει μοντέλα για τιμές του παράγοντα α εντός ενός δοθέντος διαστήματος πραγματικών αριθμών (συχνά μεταξύ 0.8 και 1.2). Επιλέγει το μοντέλο που μεγιστοποιεί την πιθανότητα των παρατηρήσεων δεδομένου των μοντέλων. Μετά την επιλογή, κανονικοποιεί τα δεδομένα με το αντίστοιχο α και ξαναεκτιμά τα ακουστικά μοντέλα. Η διαδικασία επαναλαμβάνεται μέχρι τη σύγκλιση. Με τον τρόπο αυτό, δεν υπολογίζονται οι θέσεις των συχνοτήτων φωνοσυντονισμού ή το μήκος της φωνητικής οδού, παρά μόνο ο βέλτιστος μετασχηματισμός του άξονα των συχνοτήτων (*frequency warping*), ο οποίος μπορεί να καθορίζεται και από άλλους παράγοντες (π.χ. F0).

Σημαντική συνεισφορά της VTLN τεχνικής είναι η ικανοποιητική μοντελοποίηση ομιλητών και από τα 2 φύλα, διώχνωντας την ανάγκη για ξεχωριστά μοντέλα εξαρτώμενα του φύλου (*gender-dependent*).

5.3.1 Feature-level VTLN

Το Kaldi δίνει τη δυνατότητα εφαρμογής της VTLN σε επίπεδο χαρακτηριστικών, κατά τον υπολογισμό των MFCC ή PLP χαρακτηριστικών, με δεδομένο τον επιθυμητό παράγοντα στρέβλωσης α. Η βασική λειτουργία είναι η μετακίνηση των θέσεων των κεντρικών συχνοτήτων των τριγωνικών φίλτρων σε κλίμακα mel. Η μετακίνηση αυτή γίνεται με επιβολή μιας τμηματικά γραμμικής συνάρτησης στο χώρο των συχνοτήτων. Η συνάρτηση αυτή αποτελείται από 3 γραμμικά τμήματα και απεικονίζει το διάστημα $[f_{low}$,$f_{high}]$ στο $[f_{low}^{(VTLN)}$, $f_{high}^{(VTLN)}]$. Οι συχνότητες f_{low}, f_{high} αντιστοιχούν στη χαμηλότερη και υψηλότερη συχνότητα που χρησιμοποιείται κατά τον υπολογισμό των MFCC ή PLP χαρακτηριστικών. Οι συχνότητες $f_{low}^{(VTLN)}$, $f_{high}^{(VTLN)}$ είναι συχνότητες αποκοπής της VTLN, έτσι ώστε τα τριγωνικά φίλτρα να έχουν λογικό εύρος. Ισχύει:

$$0 \leq f_{low} \leq f_{low}^{(VTLN)} < f_{high}^{(VTLN)} < f_{high} \leq f_{Nyquist} \tag{5.3.3}$$

όπου $f_{Nyquist}$ η συχνότητα Nyquist του σήματος. Στο Σχήμα 5.2 απεικονίζεται ενδεικτικά η VTLN συνάρτηση στρέβλωσης που χρησιμοποιεί το Kaldi [63].

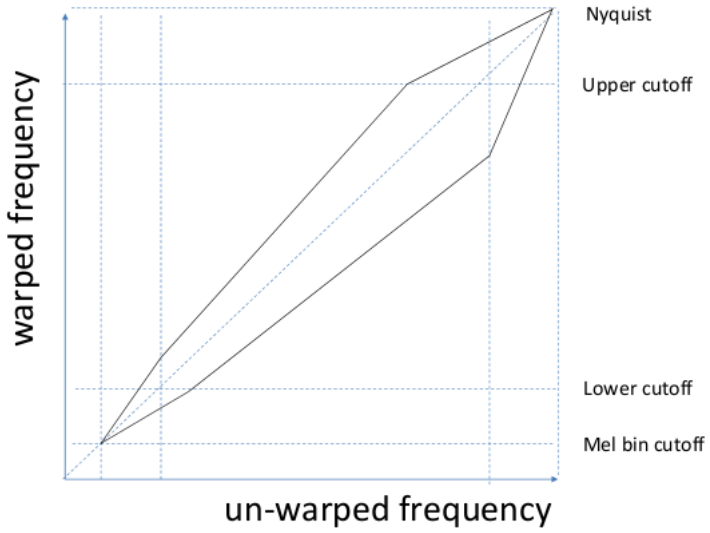
5.3.2 Linear VTLN (LVTLN)

Έχουν αναπτυχθεί διάφοροι τρόποι υλοποίησης της VTLN, όπου ο παράγοντας στρέβλωσης εκτιμάται με βάση το μοντέλο. Το Kaldi υλοποιεί τη Linear VTLN [64]. Η βασική ιδέα είναι η εφαρμογή γραμμικού μετασχηματισμού στα MFCC ή στα PLP χαρακτηριστικά, ξεχωριστό για κάθε ομιλητή, έτσι ώστε να μειωθούν οι φασματικές διαφορές μεταξύ τους. Ο μετασχηματισμός αυτός για κάθε ομιλητή εκτιμάται μεγιστοποιώντας επαναληπτικά την πιθανότητα των κανονικοποιημένων χαρακτηριστικών του, δεδομένου του μοντέλου.

Αρχικά, υπολογίζονται N γραμμικοί μετασχηματισμοί των χαρακτηριστικών των δεδομένων εκπαίδευσης, ένας για κάθε παράγοντα στρέβλωσης:

$$\mathbf{A}^{(i)}, 0 \leq i \leq N \tag{5.3.4}$$

Θέτοντας, παραδείγματος χάριν, τις τιμές των VTLN παραγόντων στρέβλωσης α στο διάστημα [0.85, 1.25] με βήμα 0.01, προκύπτουν $N = 41$ διαφορετικοί μετασχηματισμοί. Για τον υπολο-



Σχήμα 5.2: *Η VTLN συνάρτηση στρέβλωσης (τμηματικά γραμμική συνάρτηση) που χρησιμοποιεί το Kaldi. (Η γραφική απεικόνιση προέρχεται από το [63].)*

γισμό κάθε πίνακα $\mathbf{A}^{(i)}$, επιλέγεται ένα υποσύνολο των δεδομένων εκπαίδευσης, για τα οποία υπολογίζονται τα αρχικά χαρακτηριστικά $\mathbf{x}$ και τα μετασχηματισμένα $\hat{\mathbf{x}}$. Τα μετασχηματισμένα χαρακτηριστικά $\hat{\mathbf{x}}$ προκύπτουν με εφαρμογή της feature-level VTLN για το αντίστοιχο παράγοντα στρέβλωσης α (βλ. παρ. 5.3.1). Στη συνέχεια, θεωρώντας τον παρακάτω αφινικό μετασχηματισμό που απεικονίζει το $\mathbf{x}$ στο $\hat{\mathbf{x}}'$:

$$\hat{\mathbf{x}}' = \mathbf{A}'\mathbf{x} + \mathbf{b}' \tag{5.3.5}$$

επιδιώκεται ο υπολογισμός του πίνακα $\mathbf{A}'$, που θα οδηγήσει στον ζητούμενο $\mathbf{A}^{(i)}$, καθώς και ενός επιπρόσθετου όρου $\mathbf{b}'$. Ο υπολογισμός των $\mathbf{A}'$ και $\mathbf{b}'$ γίνεται με ελαχιστοποίηση του όρου:

$$\sum_k (\hat{\mathbf{x}}'_k - \hat{\mathbf{x}}_k)^\intercal (\hat{\mathbf{x}}'_k - \hat{\mathbf{x}}_k) \tag{5.3.6}$$

όπου k ο δείκτης των πλαισίων των επιλεγμένων εκφωνήσεων. Ο ζητούμενος πίνακας $\mathbf{A}^{(i)}$ προκύπτει μετά από πολλαπλασιασμό κάθε d-οστής γραμμής του πίνακα $\mathbf{A}'$ με τον όρο $\sqrt{\frac{\mathbf{\Sigma}^{(\mathbf{x})}_{d,d}}{\mathbf{\Sigma}^{(\hat{\mathbf{x}}')}_{d,d}}}$, όπου $\mathbf{\Sigma}^{(\mathbf{x})}$ ο πίνακας συνδιασποράς των χαρακτηριστικών $\mathbf{x}$ και $\mathbf{\Sigma}^{(\hat{\mathbf{x}}')}$ ο αντίστοιχος των $\hat{\mathbf{x}}'$.

Στη συνέχεια, για κάθε ομιλητή εκτιμάται ένας πίνακας μετασχηματισμού, σύμφωνα με την ακόλουθη περιγραφή. Έχοντας ένα ήδη εκπαιδευμένο μοντέλο και χρησιμοποιώντας τα αρχικά χαρακτηριστικά, υπολογίζεται ένα διάνυσμα offset $\mathbf{b}$, για κάθε πίνακα $\mathbf{A}^{(i)}$. Το διάνυσμα αυτό, προκύπτει με τη μεγιστοποίηση της αντικειμενικής συνάρτησης της CMLLR (βλ. παρ. 5.5.1) για το μετασχηματισμό:

$$\mathbf{W}^{(i)} = [\mathbf{A}^{(i)} \quad \mathbf{b}] \tag{5.3.7}$$

Το $\mathbf{W}^{(i)}$ που δίνει τη μεγαλύτερη τιμή της αντικειμενικής συνάρτησης πάνω σε όλα τα i, επιλέγεται ως ο μετασχηματισμός για τον συγκεκριμένο ομιλητή. Επισημαίνεται ότι συνέπεια του παραπάνω υπολογισμού του διανύσματος $\mathbf{b}$ (mean offset) είναι η επιπλέον εφαρμογή μιας Cepstral Mean Normalization με βάση το μοντέλο.

5.4 Maximum A-Posteriori (MAP) Adaptation

Προσαρμογή των παραμέτρων ενός ανεξάρτητου-του-ομιλητή μοντέλου μπορεί να επιτευχθεί και σύμφωνα με τη Maximum A-Posteriori (MAP) προσέγγιση. Η τεχνική αυτή λαμβάνει υπόψη την εκ-των-προτέρων (a-priori) κατανομή των παραμέτρων του μοντέλου. Χρησιμοποιώντας, δηλαδή, πρότερη γνώση, η οποία αντλείται από τα δεδομένα προσαρμογής, εκτιμάται η εκ-των-προτέρων κατανομή της πιθανότητας των παραμέτρων. Έστω $\boldsymbol{\vartheta}$ οι παράμετροι του μοντέλου και $p_0(\boldsymbol{\vartheta})$ η εκ-των-προτέρων κατανομή τους, τότε επιθυμείται να μεγιστοποιηθεί η εκ-των-υστέρων (a-posteriori) πιθανότητα, που είναι ανάλογη του γινομένου της πιθανοφάνειας επί την εκ-των-προτέρων πιθανότητα (κανόνας του Bayes), σύμφωνα με τη σχέση:

$$\hat{\boldsymbol{\vartheta}} = \arg\max_{\boldsymbol{\vartheta}} [p(D|\boldsymbol{\vartheta}) p_0(\boldsymbol{\vartheta})] \tag{5.4.1}$$

όπου D τα δεδομένα προσαρμογής. Η προσαρμογή αυτή μπορεί να γίνει ξεχωριστά για κάθε εκφώνηση ή κάθε ομιλητή κατά την αναγνώριση, ή για το σύνολο των δεδομένων αξιολόγησης. Αξίζει να σημειωθεί ότι σε περίπτωση που η εκ-των-προτέρων κατανομή είναι ομοιόμορφη, δε δίνει κάποια πληροφορία ή προτίμηση για τις παραμέτρους του μοντέλου (*non-informative prior*). Τότε, η παραπάνω σχέση οδηγεί στην κλασσική εκτίμηση του μοντέλου με βάση τη Μέγιστη Πιθανοφάνεια (Maximum Likelihood).

Θεωρώντας ένα ανεξάρτητο-του-ομιλητή μοντέλο GMM, με μέση τιμή $\boldsymbol{\mu}$ για τη Gaussian συνιστώσα m, η αντίστοιχη μέση τιμή ανανεώνεται σύμφωνα με τη σχέση:

$$\hat{\boldsymbol{\mu}} = \frac{c^{(m)}}{c^{(m)} + \tau} \bar{\boldsymbol{\mu}} + \frac{\tau}{c^{(m)} + \tau} \boldsymbol{\mu} \tag{5.4.2}$$

όπου $\bar{\boldsymbol{\mu}}$ η μέση τιμή των δεδομένων προσαρμογής, σύμφωνα με την προσέγγιση Μέγιστης Πιθανοφάνειας, $c^{(m)}$ το άθροισμα όλων των εκ-των-υστέρων πιθανοτήτων για τη Gaussian συνιστώσα m, δεδομένου των δεδομένων προσαρμογής (π.χ. ενός ομιλητή) και τ ένα βάρος, ορισμένο για την πρότερη γνώση. Όπως φαίνεται από την παραπάνω σχέση, αν η τιμή του $c^{(m)}$ είναι σχετικά μικρή, η μέση τιμή θα διατηρηθεί κοντά σε αυτή του ανεξάρτητου-του-ομιλητή μοντέλου.

Το βασικό μειονέκτημα της MAP είναι ότι επηρεάζει *τοπικά*, μόνο τις παραμέτρους των Gaussian συνιστωσών που παρατηρήθηκαν. Λειτουργεί σε επίπεδο Gaussian συνιστωσών, με αποτέλεσμα να απαιτείται μεγάλος αριθμός δεδομένων για προσαρμογή, έτσι ώστε να είναι αποτελεσματική.

5.5 Maximum Likelihood Linear Regression (MLLR)

Η Maximum Likelihood Linear Regression (MLLR) αποτελεί και αυτή μια τεχνική προσαρμογής βασιζόμενη στο μοντέλο. Η βασική ιδέα είναι η επιβολή ενός αφινικού μετασχηματισμού στις παραμέτρους του βασικού μοντέλου (συνήθως GMM ή HMM), με σκοπό την προσαρμογή του στα δεδομένα κάθε ομιλητή ξεχωριστά. Η εκτίμηση του μετασχηματισμού αυτού για κάθε ομιλητή γίνεται με μεγιστοποίηση της πιθανοφάνειας των δεδομένων του. Ο τρόπος αυτός λύνει το πρόβλημα της *τοπικότητας* της MAP (βλ. παρ. 5.4), καθώς κάθε μετασχηματισμός υπολογίζεται με βάση το σύνολο των δεδομένων του συγκεκριμένου ομιλητή.

Στην περίπτωση ενός GMM μοντέλου, με μέση τιμή $\boldsymbol{\mu}$ και διασπορά $\boldsymbol{\Sigma}$ για κάποια Gaussian συνιστώσα m, οι νέες παράμετροι ($\hat{\boldsymbol{\mu}}$ και $\hat{\boldsymbol{\Sigma}}$) δίνονται από τις παρακάτω σχέσεις:

$$\hat{\boldsymbol{\mu}} = \mathbf{W}\boldsymbol{\mu} + \mathbf{b} = \mathbf{A}\boldsymbol{\xi} \tag{5.5.1}$$

όπου $\mathbf{W} = [\mathbf{A} \quad \mathbf{b}]$ ο πίνακας μετασχηματισμού για έναν ομιλητή και $\boldsymbol{\xi} = [\boldsymbol{\mu} \quad 1]^\intercal$

$$\hat{\boldsymbol{\Sigma}} = \mathbf{L}\mathbf{B}\mathbf{L}^\intercal \quad \text{ή} \quad \hat{\boldsymbol{\Sigma}} = \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\intercal \tag{5.5.2}$$

όπου $\mathbf{L}$ ο παράγοντας Cholesky του πίνακα $\boldsymbol{\Sigma}$ ($\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\intercal$). Αν εφαρμοστεί μόνο ο μετασχηματισμός της μέσης τιμής (5.5.1), τότε προκύπτει η *mean-only* MLLR. Αν επιπλέον εφαρμοστεί και ο μετασχηματισμός της διασποράς, χρησιμοποιώντας μία από τις σχέσεις (5.5.2), τότε η τεχνική αντιστοιχεί στην κανονική MLLR.

Μια άλλη επιλογή αφορά τη μορφή του πινάκων μετασχηματισμού. Συνήθως, είναι διαγώνιοι (diagonal) ή πλήρεις (full). Επίσης, στην περίπτωση της μέσης τιμής μπορεί να προστεθεί μόνο το offset διάνυσμα $\mathbf{b}$, με μοναδιαίο $\mathbf{W}$. Προφανώς, όσο μικρότερος είναι ο αριθμός των παραμέτρων, τόσο πιο απλός είναι ο υπολογισμός του μετασχηματισμού. Το μειονέκτημα που προκύπτει είναι ότι εξαιτίας της απλότητας αυτής, μειώνεται η ικανότητα προσαρμογής και η αποτελεσματικότητα του μετασχηματισμού.

Για τον υπολογισμό των πινάκων $\mathbf{A}$ και $\mathbf{B}$ των σχέσεων (5.5.1) και (5.5.2), επιδιώκεται η μεγιστοποίηση της πιθανοφάνειας των δεδομένων παρατήρησης. Κάτι που αξίζει να σημειωθεί αφορά το κόστος υπολογισμού του πίνακα μετασχηματισμού $\mathbf{B}$ συγκριτικά με αυτό της πιθανοφάνειας των δεδομένων κατά την αναγνώριση. Η πρώτη σχέση από τις (5.5.2), ενώ δίνει γρήγορη σχετικά εκτίμηση του μετασχηματισμού της διασποράς, απαιτεί μεγάλο κόστος υπολογισμού της πιθανοφάνειας των δεδομένων για αναγνώριση [65], ειδικά όταν ο πίνακας $\hat{\boldsymbol{\Sigma}}$ είναι πλήρης, αφού:

$$\mathcal{L}(\mathbf{x}_k|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}, \mathbf{B}) = \mathcal{N}(\mathbf{x}_k; \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) \tag{5.5.3}$$

όπου $\mathbf{x}_k$ ένα δεδομένο παρατήρησης για $k = 1, ..., N$. Αντίθετα, η δεύτερη σχέση από τις (5.5.2), ενώ είναι χρονοβόρα κατά τον υπολογισμό του πίνακα $\mathbf{B}$, καθώς ακολουθείται μια επαναληπτική διαδικασία ανά γραμμή, είναι αρκετά απλή για την αναγνώριση [65]. Αυτό οφείλεται στη δυνατότητα εφαρμογής του μετασχηματισμού στα δεδομένα παρατήρησης και όχι στο μοντέλο:

$$log\mathcal{L}(\mathbf{x}_k|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}, \mathbf{B}) = log\mathcal{N}(\mathbf{B}^{-1}\mathbf{x}_k; \mathbf{B}^{-1}\hat{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) - \frac{1}{2}log|\mathbf{B}|^2 \tag{5.5.4}$$

Μια παραλλαγή της κανονικής MLLR αντιστοιχεί στις σχέσεις:

$$\hat{\boldsymbol{\mu}} = \mathbf{A}_c\boldsymbol{\mu} - \mathbf{b}_c \tag{5.5.5}$$

$$\hat{\boldsymbol{\Sigma}} = \mathbf{A}_c\boldsymbol{\Sigma}\mathbf{A}_c^\intercal \tag{5.5.6}$$

Εδώ, χρησιμοποιείται ο ίδιος πίνακας μετασχηματισμού $\mathbf{A}_c$ και για τη μέση τιμή και για τη διασπορά. Η προσέγγιση αυτή ονομάζεται Constrained MLLR (CMLLR) [65]. Και αυτή στηρίζεται στην ιδέα της εφαρμογής του μετασχηματισμού στα δεδομένα παρατήρησης με σκοπό τη βελτίωση του χρόνου αναγνώρισης. Η λογαριθμική πιθανοφάνεια προκύπτει:

$$log\mathcal{L}(\mathbf{x}_k|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}_c, \mathbf{b}_c) = log\mathcal{N}(\mathbf{A}_c^{-1}\mathbf{x}_k + \mathbf{A}_c^{-1}\mathbf{b}_c; \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \frac{1}{2}log|\mathbf{A}_c^{-1}|^2 \tag{5.5.7}$$

Δηλαδή, τα δεδομένα παρατήρησης μετασχηματίζονται σύμφωνα με τη σχέση:

$$\hat{\mathbf{x}}_k = \mathbf{A}_c^{-1}\mathbf{x}_k + \mathbf{A}_c^{-1}\mathbf{b}_c \tag{5.5.8}$$

Έτσι, η CMLLR εφαρμόζεται καθαρά ως μετασχηματισμός των χαρακτηριστικών των αντίστοιχων δεδομένων.

5.5.1 Constrained MLLR (CMLLR) ή feature-space MLLR (fMLLR)

Το Kaldi αναφέρει την Constrained MLLR και με τον όρο feature-space MLLR (fMLLR), καθώς είναι ουσιαστικά ένας αφινικός μετασχηματισμός των χαρακτηριστικών κάθε ομιλητή. Έστω ένα διάνυσμα χαρακτηριστικών $\mathbf{x}_k$ που αντιστοιχεί στο πλαίσιο k και στον ομιλητή s, τότε αυτό απεικονίζεται σύμφωνα με τον παρακάτω μετασχηματισμό:

$$\mathbf{x}_k \rightarrow \mathbf{A}^{(s)}\mathbf{x}_k + \mathbf{b}^{(s)} \tag{5.5.9}$$

Το παραπάνω μπορεί να γραφτεί και ως:

$$\mathbf{x}_k \rightarrow \mathbf{W}^{(s)}\boldsymbol{\xi}_k \tag{5.5.10}$$

όπου ο πίνακας $\mathbf{W}^{(s)} = [\mathbf{A}^{(s)} \quad \mathbf{b}^{(s)}]$ περιέχει τον τετραγωνικό $\mathbf{A}^{(s)}$ και έναν επιπρόσθετο όρο $\mathbf{b}^{(s)}$, ενώ το $\boldsymbol{\xi}_k = [\mathbf{x}_k \quad 1]^\intercal$ αντιστοιχεί στο επεκταμένο διάνυσμα χαρακτηριστικών.

Για τον υπολογισμό του πίνακα $\mathbf{W}^{(s)}$ για τον ομιλητή s, επιδιώκεται η μεγιστοποίηση μιας αντικειμενικής συνάρτησης. Παρακάτω, περιγράφεται συνοπτικά η διαδικασία υπολογισμού της συνάρτησης αυτής, για περίπτωση μοντέλου GMM με διαγώνιο πίνακα συνδιακύμανσης. Η αντικειμενική συνάρτηση ισούται με την πιθανοφάνεια των μετασχηματισμένων χαρακτηριστικών συν το λογάριθμο της ορίζουσας του πίνακα $\mathbf{A}^{(s)}$. Αν παραλειφθεί ο όρος $log(|\mathbf{A}^{(s)}|)$, η αντικειμενική συνάρτηση ισούται με [66]:

$$-0.5\sum_{m=1}^{M} c^{(sm)}\mathcal{E}\Big\{\sum_{i=1}^{d}\frac{(\mu_i^{(m)} - \mathbf{w}_i^\intercal\boldsymbol{\xi}_k)^2}{\sigma_i^{2(m)}}\Big\}^{(sm)} \tag{5.5.11}$$

όπου M αντιστοιχεί στο συνολικό αριθμό των Gaussian συνιστωσών που έχουν αντιστοιχηθεί τα δεδομένα του ομιλητή s, d αντιστοιχεί στον αριθμό των διαστάσεων των χαρακτηριστικών του, $\mathbf{w}_i^\intercal$ αντιστοιχεί στην i-οστή γραμμή του πίνακα $\mathbf{W}^{(s)}$ και με $\mathcal{E}\{.\}^{(sm)}$ συμβολίζεται η μέση τιμή για τον ομιλητή s και την Gaussian συνιστώσα m. Επίσης, $c^{(sm)} = \sum_k \gamma^{(ksm)}$, όπου με $\gamma^{(ksm)}$ συμβολίζεται η εκ-των-υστέρων Gaussian πιθανότητα για τον ομιλητή s, την Gaussian συνιστώσα m και το πλαίσιο k. Η παραπάνω ποσότητα (5.5.11) ισούται με:

$$-0.5\sum_{m=1}^{M} c^{(sm)}\sum_{i=1}^{d}\frac{\mu_i^{(m)2} - 2\mu_i^{(m)}\mathbf{w}_i^\intercal\mathcal{E}\{\boldsymbol{\xi}\}^{(sm)} + \mathbf{w}_i^\intercal\mathcal{E}\{\boldsymbol{\xi}\boldsymbol{\xi}^\intercal\}^{(sm)}\mathbf{w}_i}{\sigma_i^{2(m)}} \tag{5.5.12}$$

όπου ισχύει:

$$\mathcal{E}\{\boldsymbol{\xi}\}^{(sm)} = \begin{bmatrix}\mathcal{E}\{\mathbf{x}\}^{(sm)} \\ 1\end{bmatrix} \tag{5.5.13}$$

$$\mathcal{E}\{\boldsymbol{\xi}\boldsymbol{\xi}^{\intercal}\}^{(sm)} = \begin{bmatrix} \mathcal{E}\{\mathbf{x}\mathbf{x}^{\intercal}\}^{(sm)} & \mathcal{E}\{\mathbf{x}\}^{(sm)} \\ \mathcal{E}\{\mathbf{x}\}^{(sm)\intercal} & 1 \end{bmatrix} \tag{5.5.14}$$

Θέτοντας:

$$\mathbf{k}_i = \sum_{m=1}^{M} \frac{c^{(sm)}\mu_i^{(m)}\mathcal{E}\{\boldsymbol{\xi}\}^{(sm)}}{\sigma_i^{2(m)}} \tag{5.5.15}$$

και:

$$\mathbf{G}_i = \sum_{m=1}^{M} \frac{c^{(sm)}\mathcal{E}\{\boldsymbol{\xi}\boldsymbol{\xi}^{\intercal}\}^{(sm)}}{\sigma_i^{2(m)}} \tag{5.5.16}$$

η αντικειμενική συνάρτηση προκύπτει:

$$log(|\mathbf{A}^{(s)}|) - \sum_{i=1}^{d} \mathbf{w}_i^{\intercal}\mathbf{k}_i - 0.5\mathbf{w}_i^{\intercal}\mathbf{G}_i\mathbf{w}_i \tag{5.5.17}$$

Ο πίνακας μετασχηματισμού $\mathbf{W}^{(s)}$ εκτιμάται ακολουθώντας μια επαναληπτική διαδικασία. Ξεκινώντας με $\mathbf{A}^{(s)} = \mathbf{I}$ και $\mathbf{b}^{(s)} = 0$, ανανεώνεται κάθε γραμμή ξεχωριστά, μέχρι η αλλαγή της τιμής της αντικειμενικής συνάρτησης να είναι πολύ μικρή ή μέχρι κάποιο μέγιστο αριθμό επαναλήψεων. Η ανανέωση αυτή γίνεται σύμφωνα με τις παρακάτω σχέσεις, οι οποίες προκύπτουν ως λύση της μεγιστοποίησης της ποσότητας (5.5.17), σύμφωνα με το [66]:

$$[a, b, c] = [1, -\mathbf{c}_i^{\intercal}\mathbf{G}_i^{-1}\mathbf{k}_i, -\mathbf{c}_i^{\intercal}\mathbf{G}_i^{-1}\mathbf{c}_i] \tag{5.5.18}$$

$$f = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{5.5.19}$$

$$\mathbf{w}_i = \mathbf{G}_i^{-1}(\frac{\mathbf{c}_i}{f} + \mathbf{k}_i) \tag{5.5.20}$$

όπου $\mathbf{c}_i$ αντιστοιχεί στην i-οστή στήλη του πίνακα $\mathbf{A}^{-1(s)}$ σε κάθε επανάληψη, επεκτείνοντάς την με ένα 0 στο τέλος, έτσι ώστε να έχει διάσταση $d + 1$.

5.6 Speaker Adaptive Training (SAT)

Ιδιαίτερο ενδιαφέρον παρουσιάζει η χρησιμοποίηση κάποιας τεχνικής Προσαρμογής του Ομιλητή κατά την εκπαίδευση του μοντέλου, όπως υποδηλώνει το Speaker Adaptive Training (SAT). Αντικαθίσταται, δηλαδή, το ανεξάρτητο-του-ομιλητή μοντέλο, το οποίο θα προσαρμοζόταν απλά κατά την αξιολόγηση για τα αντίστοιχα δεδομένα. Τώρα, εκπαιδεύεται ένα μοντέλο, γνωρίζοντας τη μέθοδο προσαρμογής που θα εφαρμοστεί για την αναγνώριση, και χρησιμοποιώντας την για τα δεδομένα εκπαίδευσης. Στόχος είναι να διαχωριστούν οι παραλλαγές της φωνής που οφείλονται στη διαφορετικότητα των ομιλητών από τα καθαρά φωνητικά χαρακτηριστικά. Έτσι, ενώ η κλασσική εκπαίδευση ενός μοντέλου $\boldsymbol{\vartheta}$ με τα δεδομένα εκπαίδευσης $D^{(s)}$ κάθε ομιλητή s, για $s = 1, ..., R$, ακολουθεί τη σχέση:

$$\hat{\boldsymbol{\vartheta}} = \arg\max_{\boldsymbol{\vartheta}} \prod_{s=1}^{R} \mathcal{L}(D^{(s)}|\boldsymbol{\vartheta}) \tag{5.6.1}$$

η τεχνική εκπαίδευσης SAT μπορεί να εκφραστεί [65] σύμφωνα με τη σχέση:

$$(\hat{\boldsymbol{\vartheta}}_c, \hat{\mathcal{H}}) = \arg\max_{(\boldsymbol{\vartheta}_c, \mathcal{H})} \prod_{s=1}^{R} \mathcal{L}(D^{(s)} | \mathcal{H}^{(s)}(\boldsymbol{\vartheta}_c)) \tag{5.6.2}$$

όπου $\mathcal{H}^{(s)}$ αντιστοιχεί στο μετασχηματισμό του μοντέλου για τον ομιλητή s. Με αυτόν τον τρόπο, οι διαφορές των ομιλητών μοντελοποιούνται από το $\mathcal{H}$, ενώ τα φωνητικά χαρακτηριστικά από το συμπαγές μοντέλο $\boldsymbol{\vartheta}_c$ (*compact model*).

Στην περίπτωση εκπαίδευσης SAT με fMLLR, για παράδειγμα, ένα GMM μοντέλο εκπαιδεύεται με χρήση των μετασχηματισμένων χαρακτηριστικών. Επαναληπτικά, εκτιμάται ένας πίνακας μετασχηματισμού για κάθε διαφορετικό ομιλητή που περιλαμβάνεται στα δεδομένα εκπαίδευσης. Σκοπός είναι η μεγιστοποίηση της πιθανοφάνειας των δεδομένων του, ξεκινώντας από κάποιο βασικό GMM, εκπαιδευμένο με τα αρχικά χαρακτηριστικά. Με χρήση του πίνακα μετασχηματισμού που προκύπτει για κάθε ομιλητή, μετασχηματίζονται τα χαρακτηριστικά του. Τα μετασχηματισμένα χαρακτηριστικά του συνόλου των ομιλητών δίνονται για εκπαίδευση του νέου GMM, με Expectation Maximization. Κατά την αξιολόγηση, εφαρμόζεται κανονική fMLLR. Το μοντέλο που προέκυψε από το SAT χρησιμοποιείται για τον υπολογισμό του πίνακα μετασχηματισμού κάθε ομιλητή. Ο πίνακας αυτός επιβάλλεται στα χαρακτηριστικά του συγκεκριμένου ομιλητή με σκοπό την αναγνώριση.

Κεφάλαιο 6

Προσαρμογή του Ομιλητή για Αναγνώριση Συναισθήματος

6.1 Εισαγωγή

Το παρόν κεφάλαιο επικεντρώνεται στην εφαρμογή των κυριότερων τεχνικών για Προσαρμογή του Ομιλητή, που περιγράφηκαν στο προηγούμενο κεφάλαιο, με σκοπό την αναγνώριση συναισθήματος από φωνή. Δεδομένου ότι οι παραπάνω τεχνικές προσαρμογής έχουν δείξει σημαντική βελτίωση των συστημάτων στον τομέα της Αυτόματης Αναγνώρισης Φωνής, καθώς μειώνουν την επίδραση του διαφορετικού ομιλητή και των διαφορετικών συνθηκών ηχογράφησης, είναι σημαντικό να διερευνηθεί η συνεισφορά τους στον τομέα της Αναγνώρισης Συναισθήματος από Φωνή (Speech Emotion Recognition). Λόγω της ιδιαίτερη φύσης του συναισθήματος, η έκφρασή του ποικίλει ανάλογα με τον άνθρωπο, την κοινωνία, τον πολιτισμό. Όπως αναλύθηκε και στα Κεφάλαιο 1 και 2, παρατηρείται σημαντική διαφοροποίηση μεταξύ των ομιλητών, τόσο όσον αφορά τα χαρακτηριστικά της φωνής τους, όσο και τον τρόπο εξωτερίκευσης ενός συναισθήματος. Όμως, για ένα αυτόματο σύστημα, είναι πολύ δύσκολο, αν όχι αδύνατο, να αντιληφθεί τέτοια ιδιαίτερα χαρακτηριστικά, λαμβάνοντας ως είσοδο μόνο ένα ηχητικό σήμα. Ακόμα και η χρήση διαφορετικής γλώσσας μπορεί να επηρεάσει τη σωστή αναγνώριση του συναισθήματος, όπως επίσης και το διαφορετικό ακουστικό περιβάλλον. Συμπερασματικά, θα ήταν χρήσιμη η προσαρμογή του συστήματος σε ένα νέο ομιλητή, έτσι ώστε να είναι εύρωστο.

Μικρή προσπάθεια έχει γίνει μέχρι τώρα για την προσαρμογή ενός ομιλητή σε ένα σύστημα αναγνώρισης συναισθήματος από φωνή, εφαρμόζοντας τεχνικές που χρησιμοποιούνται κατά κόρον στην Αναγνώριση Φωνής. Ενδεικτικά, στο [67] προτείνεται μια on-line μέθοδος προσαρμογής του ομιλητή με εφαρμογή της MLLR, μετά από επιλεκτική διόρθωση των ετικετών των δεδομένων προσαρμογής, οι οποίες προκύπτουν από το ανεξάρτητο-του-ομιλητή μοντέλο. Τονίζεται ότι εξαιτίας της ασάφειας που διακρίνει τα διάφορα συναισθήματα ως προς την έκφραση και ως προς την κατανόηση τους, δημιουργούνται αρκετά λάθη επισημείωσης των δεδομένων, και γι' αυτό οι συγγραφείς αναπτύσσουν μια μέθοδο επιλεκτικής διόρθωσης των ετικετών. Στο [68] υλοποιείται η προσαρμογή ενός Universal Background Model (UBM) για κάθε εκφώνηση και η εξαγωγή του αντίστοιχου GMM υπερ-διανύσματος (διάνυσμα ορισμένο ως η αλληλουχία των μέσων τιμών των Gaussian συνιστωσών του μείγματος). Η προσαρμογή γίνεται με MAP ή MLLR, ενώ για την ταξινόμηση των GMM υπερ-διανυσμάτων χρησιμοποιείται SVM με

γραμμική συνάρτηση πυρήνα.

Εκτός από αυτές τις τεχνικές προσαρμογής του ομιλητή, έχουν προταθεί μια σειρά από μοντέλα για αναγνώριση συναισθήματος από φωνή, αρκετά πιο πολύπλοκα, με σκοπό τη μείωση της διαφοράς μεταξύ των δεδομένων εκπαίδευσης και αξιολόγησης. Ενδεικτικά, στο [54] προτείνεται η ανάθεση βαρών στα δεδομένα εκπαίδευσης ανάλογα με την ομοιότητά τους με τα δεδομένα αξιολόγησης, υλοποιώντας 3 μεθόδους εκτίμησης των βαρών αυτών. Θεωρώντας ταξινομητή SVM, το υπερεπίπεδο διαχωρισμού μετακινείται έτσι ώστε να ληφθούν υπόψη τα πιο σημαντικά δεδομένα εκπαίδευσης. Το [69] εισάγει ένα *Adaptive Denoising Autoencoder*, όπου χρησιμοποιείται πρότερη γνώση από τα δεδομένα αξιολόγησης, με σκοπό την καλύτερη εκπαίδευση του μοντέλου. Στο [70], θεωρώντας ότι οι αντίστοιχες κλάσεις των δεδομένων εκπαίδευσης και αξιολόγησης μοιράζονται κοινές εκ-των-προτέρων πιθανότητες, αναπτύσεται μία μέθοδος βασισμένη σε νευρωνικό δίκτυο 2 επιπέδων. Επίσης, στο [71], υλοποιούνται 2 προσεγγίσεις για προσαρμογή SVM μοντέλου, με χρήση μικρού αριθμού επισημειωμένων δεδομένων. Όπως μπορεί να παρατηρηθεί, οι παραπάνω τεχνικές έχουν περιορισμούς, καθώς είτε απαιτούν χρόνο για εκπαίδευση του μοντέλου γνωρίζοντας τα δεδομένα αξιολόγησης, είτε απαιτούν κάποιο αριθμό επισημειωμένων δεδομένων από αυτά.

Εδώ, παρουσιάζονται μια σειρά από πειράματα προσαρμογής των δεδομένων κάθε ομιλητή σε GMM μοντέλο, χρησιμοποιώντας τις κλασσικές τεχνικές που περιγράφηκαν στο προηγούμενο κεφάλαιο. Στόχος είναι η βελτίωση της απόδοσης του GMM ως προς την αναγνώριση της συναισθηματικής κλάσης κάθε εκφώνησης. Ως εργαλείο χρησιμοποιήθηκε το Kaldi [61] (βλ. Παράρτημα Α), τόσο για την εξαγωγή χαρακτηριστικών και την εκπαίδευση του GMM, όσο και για τις υλοποιήσεις των τεχνικών Προσαρμογής του Ομιλητή. Επίσης, δοκιμάστηκε η μετέπειτα εκπαίδευση και αξιολόγηση SVM μοντέλου, με χρήση των μετασχηματισμένων χαρακτηριστικών. Η τελευταία προσέγγιση με SVM φαίνεται να έχει τα καλύτερα αποτελέσματα.

6.2 Περιγραφή Μοντέλου

6.2.1 Δεδομένα

Ως βάση δεδομένων χρησιμοποιήθηκε το IEMOCAP (Interactive Emotional Dyadic Motion Capture database) [72]. Η συγκεκριμένη βάση αποτελείται από περίπου 12 ώρες ομιλίας (οπτικο-ακουστικά δεδομένα) και είναι χωρισμένη σε 5 συνόδους. Κάθε σύνοδος αντιστοιχεί σε μια δυαδική αλληλεπίδραση μεταξύ ενός άντρα και μιας γυναίκας, περιλαμβάνοντας τόσο γραμμένα (scripted) σενάρια όσο και αυτοσχεδιαστική επικοινωνία μεταξύ του ζεύγους. Με αυτόν τον τρόπο, τα δεδομένα είναι εμπλουτισμένα με πραγματικές καταστάσεις και συναισθήματα, εκτός από την προσποιούμενη συνομιλία μεταξύ δύο ηθοποιών. Τα πραγματικά συναισθήματα είναι και αυτά που ποικίλουν πιο πολύ ως προς την έκφραση τους, και έτσι καθιστούν πιο δύσκολη την αναγνώρισή τους. Έτσι, η ύπαρξη αυτοσχεδιασμού σε αυτή τη βάση δεδομένων συνιστά σημαντικό χαρακτηριστικό της και τη διαφοροποιεί από άλλες βάσεις, οι οποίες περιλαμβάνουν αποκλειστικά απαγγελία κειμένου από ηθοποιούς με ταυτόχρονη θεατρική έκφραση του αντίστοιχου συναισθήματος.

Κάθε διάλογος της συλλογής αυτής έχει χωρισθεί χειροκίνητα σε εκφωνήσεις, με μέση διάρκεια περίπου 4.5 δευτερόλεπτα (από 1 έως 8 δευτερόλεπτα περίπου το καθένα). Κάθε εκφώνηση έχει επισημειωθεί από 6 ανθρώπους-επισημειωτές, τόσο με διακριτές ετικέτες (χαρά,

λύπη, θυμός, ουδέτερο κ.λ.π.), όσο και με ακέραιες τιμές για τους συναισθηματικούς άξονες Valence, Activation και Dominance. Στο παρόν κεφάλαιο, χρησιμοποιούνται μόνο διακριτές ετικέτες και μόνο οι ακουστικές εκφωνήσεις, στις οποίες μπορεί να αντιστοιχηθεί μία ετικέτα με βάση την πλειοψηφία. Συγκεκριμένα, θεωρούνται 4 κλάσεις: χαρά, λύπη, θυμός και το ουδέτερο, όπου στην κλάση χαρά περιλαμβάνονται τόσο οι επισημειωμένες εκφωνήσεις με την ετικέτα χαρά όσο και αυτές με την ετικέτα ενθουσιασμός. Ο συνολικός αριθμός δεδομένων προκύπτει 5531 εκφωνήσεις, οι οποίες χωρίζονται αναλυτικά στις παρακάτω συναισθηματικές κλάσεις:

- *χαρά*: χαρά (595 εκφωνήσεις), ενθουσιασμός (1041 εκφωνήσεις)
- *λύπη*: λύπη (1084 εκφωνήσεις)
- *θυμός*: θυμός (1103 εκφωνήσεις)
- *ουδέτερο*: ουδέτερο (1708 εκφωνήσεις)

Όσον αφορά το χωρισμό των παραπάνω δεδομένων για εκπαίδευση και αξιολόγηση, ακολουθείται *10-fold leave-one-speaker-out* προσέγγιση. Δηλαδή, οι εκφωνήσεις κάθε ενός από τους 10 διαφορετικούς ομιλητές χρησιμοποιούνται για αξιολόγηση του μοντέλου, το οποίο έχει εκπαιδευτεί με τα συνολικά δεδομένα των άλλων 9 ομιλητών. Τα τελικά αποτελέσματα που παρουσιάζονται προκύπτουν ως η μέση τιμή του ποσοστού επιτυχίας στους 10 ομιλητές αξιολόγησης.

6.2.2 Ακουστικά Χαρακτηριστικά

Για την εξαγωγή των ακουστικών χαρακτηριστικών, κάθε εκφώνηση χωρίστηκε σε πλαίσια διάρκειας 25 msec, με μετακίνηση του πλαισίου κατά 10 msec κάθε φορά. Ως ακουστικά χαρακτηριστικά για κάθε πλαίσιο, εξάχθηκαν η Ενέργεια και 12 MFCCs (συνιστώσες 1-12). Με τον όρο Ενέργεια αναφέρεται η υπολογισμένη ενέργεια του κάθε πλαισίου πριν οποιαδήποτε επεξεργασία του, σε λογαριθμική κλίμακα. Για τον υπολογισμό των MFCCs, χρησιμοποιήθηκαν 23 τριγωνικά φίλτρα. Τα παραπάνω αποτελούν τις προκαθορισμένες επιλογές κατά τον υπολογισμό των χαρακτηριστικών με το εργαλείο Kaldi.

Μετά την εξαγωγή των 13 παραπάνω ακουστικών χαρακτηριστικών, υπολογίζονται η μέση τιμή και η διακύμανση κάθε χαρακτηριστικού, κατά μήκος όλων των πλαισίων ενός ομιλητή ή μιας εκφώνησης. Με χρήση αυτών των στατιστικών, μπορεί να εφαρμοστεί CMN ή CMVN στα ακουστικά χαρακτηριστικά. Επιπρόσθετα, καταχωρούνται η πρώτη και η δεύτερη παράγωγος των κανονικοποιημένων χαρακτηριστικών, με αποτέλεσμα το τελικό διάνυσμα για κάθε πλαίσιο να έχει 39 στοιχεία.

6.2.3 Μοντέλο Μείγματος Gaussian Συνιστωσών

Το Kaldi, ως εργαλείο αναγνώρισης φωνής, υλοποιεί HMM μοντέλα (Hidden Markov Models). Συνοπτικά, ένα HMM αντιστοιχεί σε μία αλυσίδα καταστάσεων, κάθε μία από τις οποίες παράγει μία παρατήρηση. Η ακολουθία των παρατηρήσεων θεωρείται γνωστή, ενώ η ακολουθία των καταστάσεων άγνωστη (κρυμμένη). Τέτοια μοντέλα έχουν αποδειχτεί πολύ χρήσιμα σε εφαρμογές αναγνώρισης φωνής, όπου μία εκφώνηση είναι ουσιαστικά μια ακολουθία φωνημάτων.

Θεωρώντας ότι μία κατάσταση αντιστοιχεί σε ένα φώνημα, η πιθανοφάνεια ενός διανύσματος χαρακτηριστικών (παρατήρηση) δεδομένου του συγκεκριμένου φωνήματος υπολογίζεται με βάση Μοντέλο Μείγματος Gaussian Συνιστωσών (GMM).

Στην περίπτωση, όμως, εφαρμογών αναγνώρισης συναισθήματος από φωνή, μία εκφώνηση αντιστοιχείται σε μία ετικέτα. Έτσι, το αντίστοιχο μοντέλο θα πρέπει να αποτελείται από μία μόνο κατάσταση. Με αυτόν τον τρόπο, το HMM απλοποιείται σε GMM. Στον παρόν κεφάλαιο, υλοποιούνται 4 μοντέλα GMM, ένα για κάθε μία από τις ορισμένες συναισθηματικές κλάσεις. Κάθε GMM έχει διαγώνιο πίνακα συνδιακύμανσης.

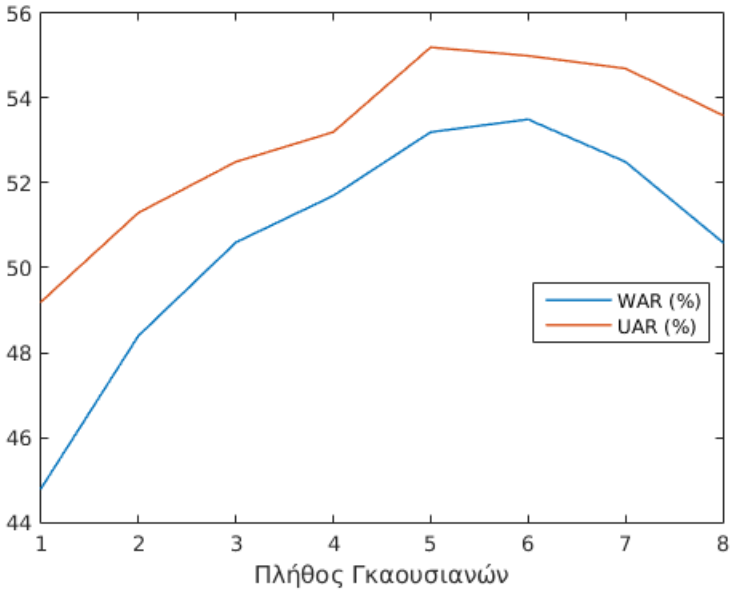
6.2.4 Αξιολόγηση Μοντέλου

Ως μετρικές για την αξιολόγηση του μοντέλου, χρησιμοποιήθηκαν οι: Weighted Average Recall (WAR) και Unweighted Average Recall (UAR). Η μετρική WAR αντιστοιχεί στο ποσοστό επιτυχίας (accuracy) επί του συνολικού αριθμού των εκφωνήσεων. Η μετρική UAR ορίζεται ως η μέση τιμή των ανακλήσεων (recall) των συναισθηματικών κλάσεων. Με τον όρο ανάκληση για μία κλάση θεωρείται το ποσοστό των σωστά ταξινομημένων εκφωνήσεων σε αυτή την κλάση επί των συνολικών εκφωνήσεων της συγκεκριμένης κλάσης. Αξίζει να σημειωθεί ότι, αρκετές φορες, η δεύτερη μετρική (UAR) δίνει καλύτερη εικόνα της επίδοσης του συστήματος, καθώς την εκφράζει ανεξάρτητα από τον αριθμό των διαθέσιμων δεδομένων ανά κλάση. Η τιμή της δεν θα καθοριστεί αναγκαστικά από το ποσοστό επιτυχίας της πολυπληθέστερης κλάσης.

Όπως αναφέρθηκε και παραπάνω, ακολουθείται *10-fold leave-one-speaker-out* προσέγγιση. Έτσι, κάθε μετρική όπου αναγράφεται, αποτελεί μέση τιμή των 10 αντίστοιχων μετρικών που προκύπτουν στα 10 πειράματα. Σε κάθε πείραμα, ως δεδομένα αξιολόγησης θεωρούνται οι συνολικές εκφωνήσεις ενός από τους 10 ομιλητές.

6.3 Πλήθος Συνιστωσών ανά GMM

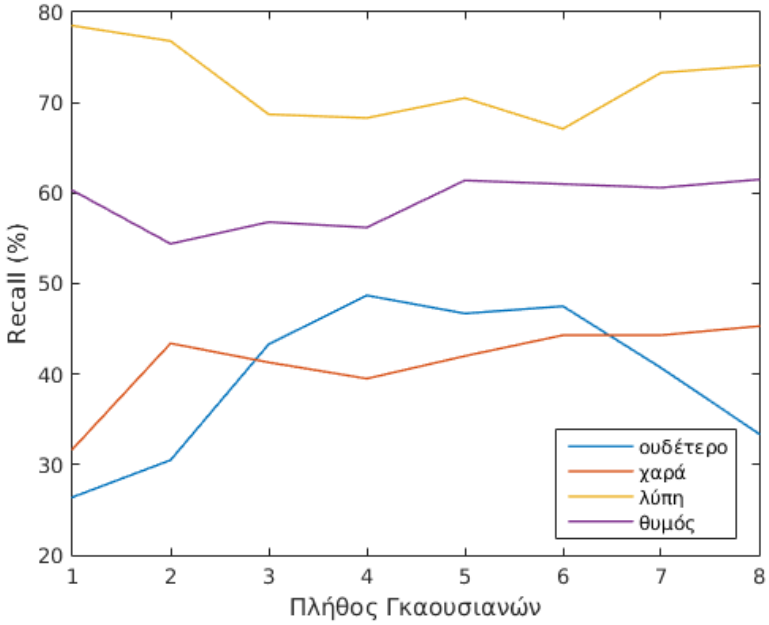
Μία σημαντική παράμετρος κάθε μοντέλου GMM αφορά το πλήθος των Gaussian συνιστωσών του. Στο κεφάλαιο 4, το μοντέλο που χρησιμοποιήθηκε για τα πειράματα κανονικοποίησης, αποτελείται από μόνο μία Gaussian συνιστώσα. Εδώ, σκοπός είναι η επιλογή του αριθμού των συνιστωσών ανά μοντέλο, έτσι ώστε να προκύψει το καλύτερο δυνατό αποτέλεσμα με τα συγκεκριμένα ακουστικά χαρακτηριστικά (βλ. παρ. 6.2.2), όπου έχει εφαρμοστεί CMN ανά ομιλητή. Σημειώνοντας, για την εκπαίδευση του GMM πραγματοποιούνται 15 επαναλήψεις. Ξεκινώντας με 1 συνιστώσα ανά μοντέλο, πραγματοποιείται αύξηση κατά 1 σε κάθε επανάληψη, μέχρι κάποιο προκαθορισμένο πλήθος Gaussian συνιστωσών ανά μοντέλο. Στο παρακάτω διάγραμμα (Σχήμα 6.1) απεικονίζονται τα αποτελέσματα ως προς τον αριθμό των Gaussian συνιστωσών:



Σχήμα 6.1: *Μετρικές WAR (%) και UAR (%) ως προς τον αριθμό των Gaussian συνιστωσών ανά μοντέλο GMM.*

Σύμφωνα με το παραπάνω σχήμα, παρατηρείται αύξηση της απόδοσης του συστήματος με την αύξηση του πλήθους των Gaussian συνιστωσών, τουλάχιστον μέχρι κάποιον αριθμό (μέχρι περίπου τις 5-6 συνιστώσες ανά GMM). Ακολουθεί σταδιακή πτώση των ποσοστών. Συμπερασματικά, μπορούν να επιλεχθούν 5 Gaussian συνιστώσες ανά GMM, οι οποίες αντιστοιχούν στο υψηλότερο ποσοστό UAR (WAR 53.2% και UAR 55.2%). Επιλέχθηκε η μετρική UAR, καθώς δίνει πιο αντιπροσωπευτικό αποτέλεσμα του συστήματος, όταν οι κλάσεις είναι μη ισορροπημένες (ελαφρώς διαφορετικός αριθμός δεδομένων ανά κλάση).

Ιδιαίτερο ενδιαφέρον, στη συγκεκριμένη περίπτωση, παρουσιάζει η απεικόνιση της ανάκλησης (recall) για κάθε συναισθηματική κλάση, ως προς το πλήθος των Gaussian συνιστωσών ανά μοντέλο (βλ. Σχήμα 6.2). Παρατηρείται ότι η ανάκληση της κλάσης *λύπη* είναι σημαντικά υψηλότερη από αυτή των άλλων κλάσεων. Γενικά, το μοντέλο, δείχνει να έχει ορθή αναγνώριση των εκφωνήσεων με ετικέτα λύπη σε πολύ μεγάλο ποσοστό (κοντά στο 80%), εις βάρος των άλλων κλάσεων. Αξίζει να σημειωθεί, επίσης, η εξισορρόπηση των ανακλήσεων με αύξηση του πλήθους των Gaussian συνιστωσών περίπου στις 5 συνιστώσες, καθώς οι χαμηλότερες ανακλήσεις αυξάνονται και οι υψηλότερες μειώνονται. Με αυτόν τον τρόπο, αποκτάται μια πιο δίκαιη αντιμετώπιση του μοντέλου απέναντι στις διαφορετικές κλάσεις.



Σχήμα 6.2: *Γραφική απεικόνιση της ανάκλησης (recall) για κάθε συναισθηματική κλάση ως προς τον αριθμό των Gaussian συνιστωσών ανά μοντέλο GMM.*

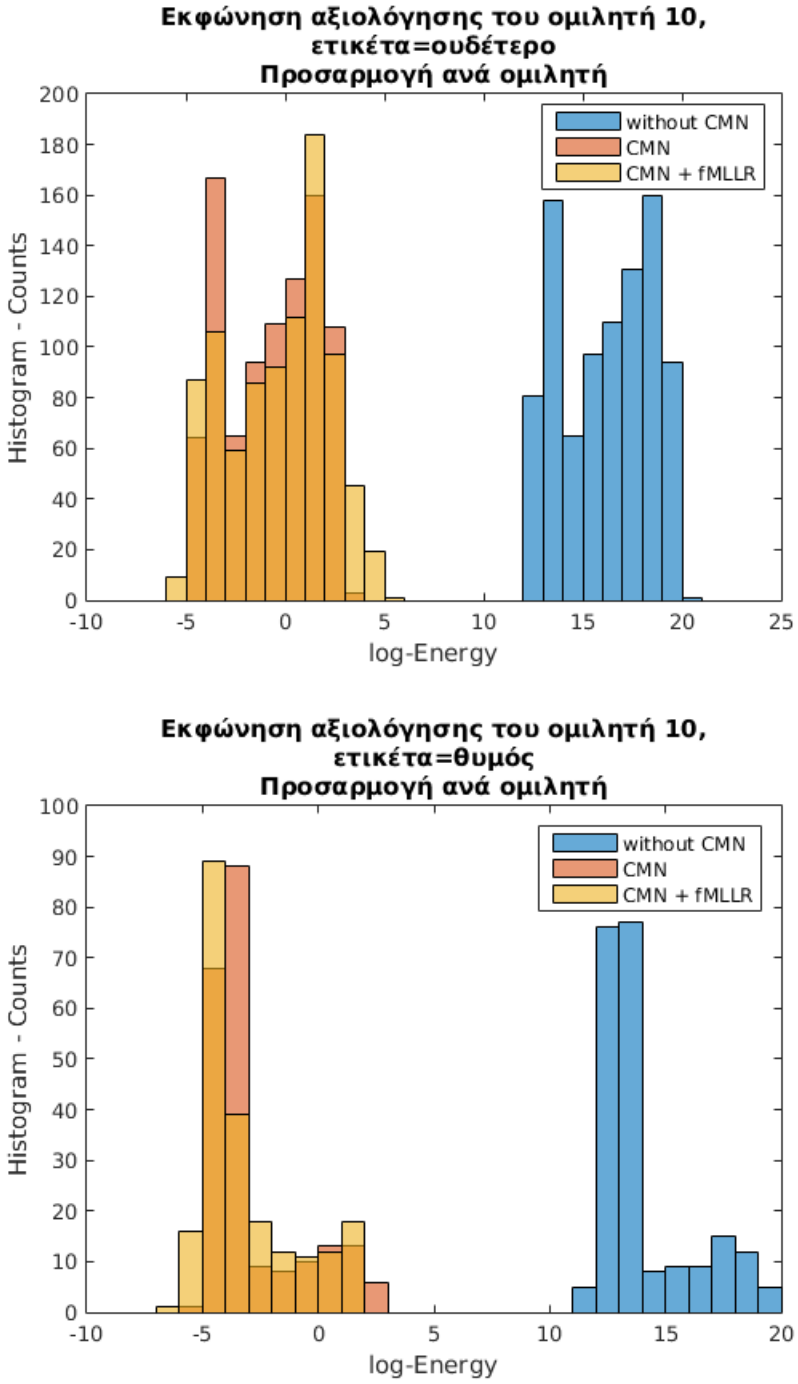
6.4 Προσαρμογή του Ομιλητή

Με δεδομένο το Μοντέλο Μείγματος 5 Gaussian Συνιστωσών, σύμφωνα με τα παραπάνω, εφαρμόζονται οι τεχνικές Προσαρμογής του Ομιλητή που παρουσιάστηκαν στο προηγούμενο κεφάλαιο. Στόχος είναι η βελτίωση της απόδοσης του GMM ως προς την αναγνώριση της συναισθηματικής κλάσης κάθε εκφώνησης. Επιθυμητή, επίσης, είναι η αξιολόγηση της προσφοράς της κάθε τεχνικής από αυτές στον τομέα της Αναγνώρισης Συναισθήματος από Φωνή. Παρακάτω, αναγράφονται τα αποτελέσματα για κάθε τεχνική, όπου τόσο η CMN/CMVN όσο και οι υπόλοιπες τεχνικές προσαρμογής, έχουν εφαρμοστεί ανά ομιλητή. Παρατηρώντας ότι η κανονικοποίηση των χαρακτηριστικών με χρήση της CMN δίνει μεγαλύτερο ποσοστό επιτυχίας, συγκριτικά με τη CMVN ή χωρίς καθόλου κανονικοποίηση, οι υπόλοιπες τεχνικές εφαρμόστηκαν μετά από CMN.

	WAR (%)	UAR (%)
χωρίς CMN	51.5	53.4
με CMN	53.2	55.2
με CMVN	52.0	54.8
CMN + Linear VTLN	52.4	53.5
CMN + MAP	**53.6**	55.2
CMN + fMLLR	51.2	54.7
CMN + SAT - fMLLR	53.4	**56.3**

Πίνακας 6.1: *Μετρικές WAR (%) και UAR (%) για κάθε τεχνική Προσαρμογής του Ομιλητή σε μοντέλο GMM με 5 Gaussian συνιστώσες.*

Το κύριο συμπέρασμα αφορά την υπεροχή της τεχνικής εκπαίδευσης SAT, όπου εδώ υλοποιήθηκε με χρήση της τεχνικής προσαρμογής fMLLR (βλ. παρ. 5.5.1) και πλήρη πίνακα μετασχηματισμού. Σύμφωνα με τη μετρική UAR, φαίνεται ότι ακόμα και η απλή εφαρμογή fMLLR μόνο κατά την αξιολόγηση δίνει σύστημα αρκετά εύρωστο, σε σύγκριση με τις άλλες



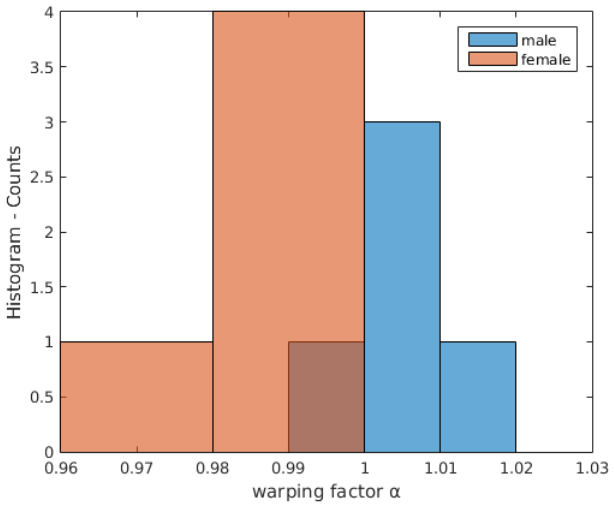
Σχήμα 6.3: *Ιστογράμματα της Ενέργειας: χωρίς CMN (μπλε), με CMN (κόκκινο) και με CMN και fMLLR (κίτρινο) για 2 εκφωνήσεις αξιολόγησης του ομιλητή 10, με ετικέτα ουδέτερο (πάνω) και θυμός (κάτω).*

τεχνικές Προσαρμογής του Ομιλητή. Επιπρόσθετα, ενώ και η Linear VTLN και η MAP απαιτούν επιπλέον εκπαίδευση του μοντέλου, η εκπαίδευση με SAT βελτιώνει σημαντικά την απόδοση του συστήματος αναγνώρισης συναισθήματος. Συμπερασματικά, η συγκεκριμένη τεχνική προσαρμοζόμενης εκπαίδευσης διαχωρίζει με μεγαλύτερη επιτυχία τις παραλλαγές της φωνής που οφείλονται στη διαφορετικότητα των ομιλητών από τα καθαρά φωνητικά χαρακτηριστικά.

Στο Σχήμα 6.3, απεικονίζονται τα ιστογράμματα της Ενέργειας για δύο εκφωνήσεις αξιολόγησης του δέκατου ομιλητή (άντρας), με ετικέτα *ουδέτερο* και *θυμός* αντίστοιχα. Περιλαμβάνεται η Ενέργεια πριν οποιαδήποτε κανονικοποίηση (χωρίς CMN), μετά την εφαρμογή CMN και μετά την εφαρμογή fMLLR για αξιολόγηση του συστήματος SAT. Παρατηρούμε ότι η εφαρμογή CMN οδηγεί σε μετακίνηση των τιμών κοντά στο 0, όπως είναι λογικό, και χωρίς ιδιαίτερη αλλαγή του σχήματος του ιστογράμματος. Ενδιαφέρον παρουσιάζει η μετακίνηση των τιμών που προκαλεί η τεχνική fMLLR. Μετά την προσαρμογή, η log-Ενέργεια της ουδέτερης εκφώνησης παίρνει περισσότερες θετικές τιμές, ενώ η log-Ενέργεια της θυμωμένης εκφώνησης μετακινείται προς πιο αρνητικές τιμές. Αποτέλεσμα είναι η διαφοροποίηση και η ορθότερη αναγνώριση των συναισθηματικών κλάσεων.

Τέλος, πρέπει να τονιστεί ότι η MAP ως τεχνική είναι αρκετά πιο γρήγορη από τη VTLN και τη SAT, με αποτέλεσμα να κρίνεται καταλληλότερη σε περίπτωση συστήματος όπου ενδιαφέρει η ταχύτητά του. Προϋπόθεση, όμως, είναι η ύπαρξη μεγάλου αριθμού δεδομένων για κάθε ομιλητή, έτσι ώστε το προσαρμοσμένο σύστημα να είναι εύρωστο. Στη συγκεκριμένη περίπτωση, παρατηρείται ότι οι 500-600 περίπου διαθέσιμες εκφωνήσεις ανά ομιλητή δίνουν μικρή βελτίωση του ποσοστού επιτυχίας σε σχέση με το απλό GMM.

Ενδεικτικά, παρακάτω απεικονίζεται το ιστόγραμμα των VTLN παραγόντων στρέβλωσης για τους 10 ομιλητές αξιολόγησης, όταν εφαρμόζεται Linear VTLN (βλ. Σχήμα 6.4). Οι ομιλητές χωρίστηκαν ανάλογα με το φύλο τους, με σκοπό να φανεί ο διαχωρισμός που αντιλαμβάνεται η τεχνική VTLN. Όπως φαίνεται και από το σχήμα, η τεχνική αυτή επιτυγχάνει ικανοποιητική μοντελοποίηση του ομιλητή με βάση το φύλο. Γενικά, ο παράγοντας στρέβλωσης των γυναικών τείνει να είναι μικρότερος από αυτόν των αντρών.



Σχήμα 6.4: *Ιστόγραμμα του παράγοντα στρέβλωσης α, κατά την εφαρμογή της Linear VTLN, για τους 10 ομιλητές αξιολόγησης.*

6.4.1 Σύγκριση Μορφής Πινάκων Μετασχηματισμού

Θεωρώντας ως πιο εύρωστο σύστημα το αποτέλεσμα εφαρμογής SAT με fMLLR, σύμφωνα με τα παραπάνω, συγκρίνεται η μορφή των πινάκων μετασχηματισμού fMLLR σε αυτήν την περίπτωση. Όπως αναφέρθηκε και στην παράγραφο 5.5, ο πίνακας μετασχηματισμού κάθε ομιλητή μπορεί να είναι πλήρης ή διαγώνιος, ή ακόμα και μοναδιαίος με επιπρόσθετη τελευταία στήλη (offset). Προφανώς, όσο λιγότερες είναι οι παράμετροι του πίνακα αυτού, τόσο μικρότερο είναι και το κόστος υπολογισμού. Όμως, σε περίπτωση όπου είναι διαθέσιμα αρκετά δεδομένα για έναν ομιλητή, ο πλήρης πίνακας θα οδηγήσει πιθανότατα σε καλύτερη προσαρμογή του συστήματος. Στον Πίνακα 6.2 παρουσιάζονται τα αποτελέσματα για τις 3 μορφές πίνακα fMLLR, όταν εφαρμόζεται εκπαίδευση SAT.

	πλήρης	διαγώνιος	offset
WAR (%)	**53.4**	51.7	51.0
UAR (%)	**56.3**	54.2	53.9

Πίνακας 6.2: *Σύγκριση μορφής πινάκων μετασχηματισμού fMLLR, κατά την εφαρμογή εκπαίδευσης SAT με fMLLR.*

Σημειώνοντας ότι για κάθε ομιλητή είναι διαθέσιμες περίπου 500-600 εκφωνήσεις, παρατηρείται ότι είναι αρκετές για την προσαρμογή του συστήματος, με αποτέλεσμα να λειτουργεί καλύτερα η χρήση πλήρη πίνακα μετασχηματισμού.

6.5 Προσαρμογή ανά Ομιλητή ή Εκφώνηση

Όλες οι παραπάνω τεχνικές προσαρμογής εφαρμόστηκαν ανά ομιλητή, σύμφωνα δηλαδή με την κλασσική υλοποίησή τους στον τομέα της Αναγνώρισης Φωνής. Ουσιαστικά, επιδιώκουν την ελαχιστοποίηση της διαφορετικότητας μεταξύ των ομιλητών ή των συνθηκών ηχογράφησης, έτσι ώστε να επιτευχθεί καλύτερη αναγνώριση της ομιλίας κάθε νέου ομιλητή. Επιβάλλοντας, όμως, ένα καθολικό μετασχηματισμό ανά ομιλητή, δε λαμβάνεται υπόψη η διαφορετική έκφραση κάθε συναισθήματος από κάθε ομιλητή. Όπως αναφέρθηκε και στην εισαγωγή, η ιδιαίτερη φύση του συναισθήματος καθιστά δύσκολη την αναγνώρισή του, καθώς είναι εφικτή η δημιουργία ασάφειας. Έτσι, κρίνεται απαραίτητη η διερεύνηση της καθολικότητας της προσαρμογής, αλλά και η ξεχωριστή αντιμετώπιση κάθε συναισθηματικού μοντέλου.

Σε αυτήν την κατεύθυνση, πραγματοποιήθηκαν τα πειράματα που παρουσιάζονται στον παρακάτω πίνακα (Πίνακας 6.3) με βάση τη μετρική UAR (%). Τα πειράματα αυτά, αναφορικά με τις γραμμές αφορούν: απλό GMM (μείγμα 5 Gaussian συνιστωσών) με εφαρμογή CMN στα χαρακτηριστικά, καθώς και τις τρεις βασικές τεχνικές προσαρμογής (VTLN, MAP και SAT με fMLLR) πάνω σε χαρακτηριστικά κανονικοποιημένα με CMN. Τόσο η CMN, όσο και η προσαρμογή με τις τρεις αυτές τεχνικές μπορεί να εφαρμοστεί ανά ομιλητή ή ανά εκφώνηση. Οι τρεις πρώτες στήλες παρουσιάζουν αυτή την επιλογή συνδυαστικά, από τις οποίες η πρώτη στήλη αντιστοιχεί στην κλασσική περίπτωση της προηγούμενης παραγράφου. Η τελευταία στήλη διαφοροποιείται, όπου η προσαρμογή γίνεται με χρήση του συνόλου των εκφωνήσεων ενός ομιλητή που αντιστοιχούν μόνο σε μια συγκεκριμένη συναισθηματική κλάση. Με αυτόν

τον τρόπο, αντιμετωπίζεται κάθε συναίσθημα κάθε ομιλητή ξεχωριστά, λαμβάνοντας υπόψη την ιδιαίτερη συναισθηματική έκφραση κάθε ατόμου.

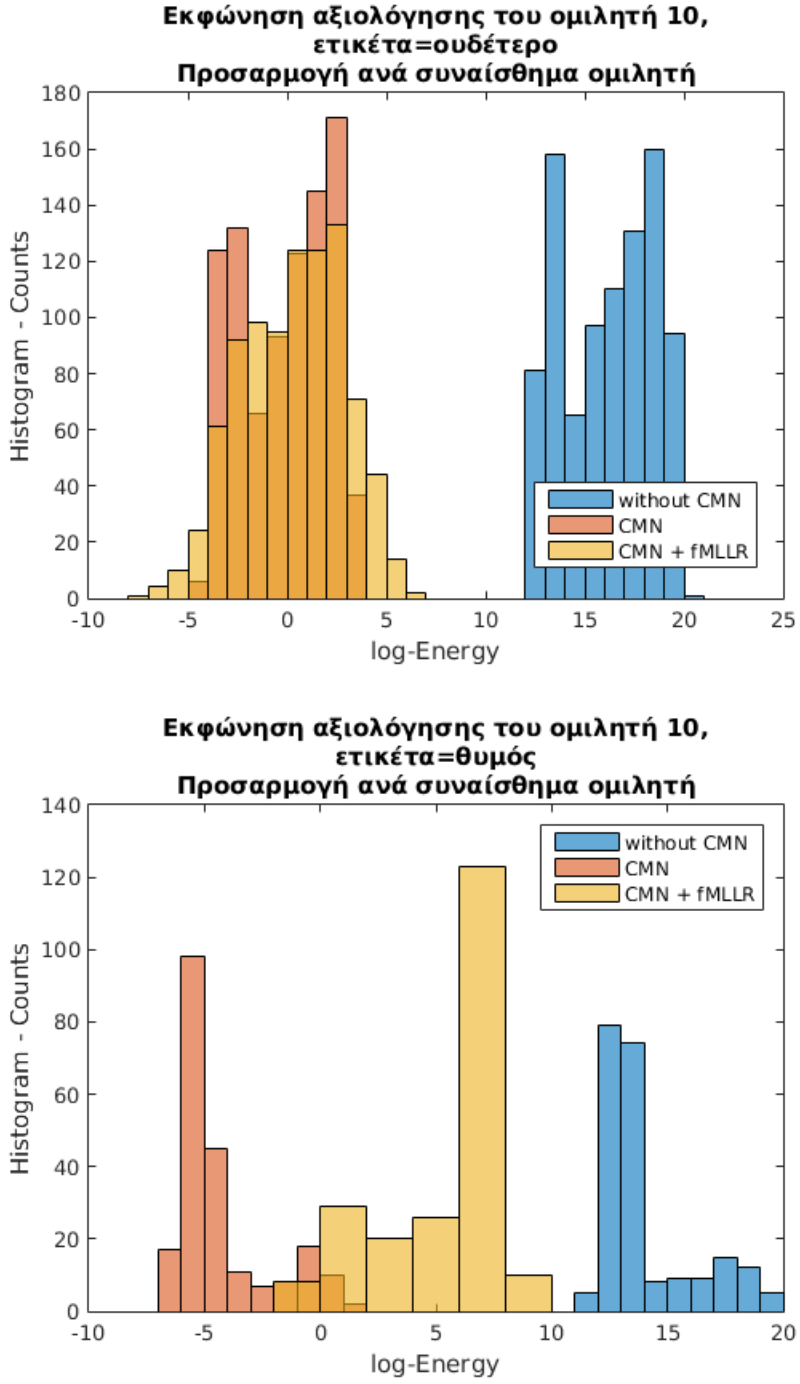
Επεξηγηματικά για την τελευταία στήλη του Πίνακα 6.3, έχει θεωρηθεί γνωστή η συναισθηματική κλάση κάθε εκφώνησης κατά την προσαρμογή του μοντέλου, τόσο κατά την εκπαίδευση όσο και κατά την αξιολόγηση, έτσι ώστε να χωριστούν τα δεδομένα κατάλληλα. Η υπόθεση αυτή γίνεται, με σκοπό να πραγματοποιηθεί ξεχωριστή αντιμετώπιση κάθε συναισθηματικής κλάσης κάθε ομιλητή και να ληφθούν υπόψη τα ιδιαίτερα στοιχεία προσωπικής έκφρασης. Με αυτόν τον τρόπο, αξιολογείται η λειτουργικότητα των τεχνικών προσαρμογής, όταν εφαρμόζονται ξεχωριστά για κάθε κλάση. Όμως, σε πραγματικές εφαρμογές δεν είναι δυνατή η εκ-των-προτέρων γνώση της κλάσης κάθε εκφώνησης. Προφανώς, κατά την αξιολόγηση του συγκεκριμένου μοντέλου, οι ετικέτες των δεδομένων δε θεωρήθηκαν γνωστές.

	CMN → ανά ομιλητή	CMN → ανά ομιλητή	CMN → ανά εκφώνηση	CMN → ανά συναίσθημα ομιλητή
	Προσαρμογή → ανά ομιλητή	Προσαρμογή → ανά εκφώνηση	Προσαρμογή → ανά εκφώνηση	Προσαρμογή → ανά συναίσθημα ομιλητή
CMN	55.2	55.2	50.7	63.4
CMN + Linear VTLN	53.5	54.2	51.8	65.8
CMN + MAP	55.2	55.2	50.6	59.7
CMN + SAT - fMLLR	**56.3**	54.1	51.5	**65.9**

Πίνακας 6.3: *Μετρική UAR (%) για 4 περιπτώσεις εφαρμογής των τεχνικών προσαρμογής στο μοντέλο GMM (μείγμα 5 Gaussian συνιστωσών). Ο όρος Προσαρμογή αναφέρεται σε μία από τις τεχνικές Linear VTLN, MAP και SAT με fMLLR, ανάλογα τη γραμμή.*

Σύμφωνα με την πρώτη γραμμή του παραπάνω πίνακα, όπου στα ακουστικά χαρακτηριστικά επιβάλλεται μόνο CMN, η κανονικοποίηση τους είναι αποτελεσματικότερη όταν εφαρμόζεται ανά ομιλητή, συγκριτικά με την εφαρμογή της ανά εκφώνηση. Το ίδιο παρατηρείται, ως επί το πλείστον, και στις υπόλοιπες τεχνικές προσαρμογής. Συμπεραίνεται ότι τα διαθέσιμα δεδομένα ανά ομιλητή είναι αρκετά για την αποτελεσματική προσαρμογή καθενός από αυτούς. Αντίθετα, στην προσαρμογή ανά εκφώνηση, διατίθεται μόνο μία εκφώνηση για προσαρμογή στο μοντέλο, με αποτέλεσμα να μην είναι αρκετή. Επιπλέον, σε αυτή την περίπτωση το κόστος υπολογισμού αυξάνεται κατακόρυφα, καθώς κάθε τεχνική εφαρμόζεται πολλαπλάσιες φορές, μία φορά για κάθε εκφώνηση. Ωστόσο, παρατηρείται μια μικρή αύξηση του ποσοστού επιτυχίας στην περίπτωση της προσαρμογής με Linear VTLN, όταν αυτή πραγματοποιείται ανά εκφώνηση και η CMN ανά ομιλητή. Όσον αφορά την τεχνική MAP, δε δείχνει να προσφέρει στο συγκεκριμένο πρόβλημα προσαρμογής, συγκριτικά με το απλό GMM.

Αξιολογώντας την περίπτωση της τελευταίας στήλης, όπου η προσαρμογή έχει γίνει ανά συναισθηματική κλάση κάθε ομιλητή, παρατηρείται αξιοσημείωτη βελτίωση του ποσοστού επιτυχίας (κοντά στο 9-10%), συγκριτικά με την κλασσική εφαρμογή κάθε τεχνικής καθολικά ανά ομιλητή. Η παρατήρηση αυτή ενισχύει την πεποίθηση της διαφορετικότητας της συναισθηματικής έκφρασης μεταξύ των ομιλητών και δείχνει τη σημασία ξεχωριστής αντιμετώπισης κάθε συναισθηματικής κλάσης κάθε ομιλητή.



Σχήμα 6.5: *Ιστογράμματα της Ενέργειας: χωρίς CMN (μπλε), με CMN (κόκκινο) και με CMN και fMLLR (κίτρινο) για 2 εκφωνήσεις αξιολόγησης του ομιλητή 10, με ετικέτα ουδέτερο (πάνω) και θυμός (κάτω). Οι τεχνικές προσαρμογής έχουν εφαρμοστεί ανά συναισθηματική κλάση ομιλητή.*

Επιλέγοντας δύο εκφωνήσεις αξιολόγησης του δέκατου ομιλητή (τις ίδιες με το Σχήμα 6.3), παρουσιάζονται τα ιστογράμματα της Ενέργειάς τους πριν οποιαδήποτε κανονικοποίηση (χωρίς CMN), μετά από εφαρμογή CMN και μετά από εφαρμογή fMLLR για αξιολόγηση του

συστήματος SAT (βλ. Σχήμα 6.5). Οι παραπάνω τεχνικές προσαρμογής, που απεικονίζονται, έχουν εφαρμοστεί ανά συναισθηματική κλάση του ομιλητή. Από τα δύο σχήματα, το πάνω αντιστοιχεί σε εκφώνηση της κλάσης *ουδέτερο*, ενώ το κάτω σε εκφώνηση της κλάσης *θυμός*. Και στις δύο περιπτώσεις, παρατηρείται μεταβολή των τιμών της Ενέργειας κατά την εφαρμογή της CMN προς το 0, όπως είναι άλλωστε λογικό. Η κύρια διαφορά εντοπίζεται κατά την εφαρμογή της fMLLR, όπου η Ενέργεια της ουδέτερης εκφώνησης διατηρείται σε παρόμοιες τιμές, ενώ η Ενέργεια της θυμωμένης εκφώνησης μετακινείται σε πιο υψηλές τιμές. Η μετακίνηση αυτή διαφοροποιεί την τελική ενέργεια των εκφωνήσεων που αντιστοιχούν στην κλάση θυμός από τις ουδέτερες, με αποτέλεσμα την ορθότερη αναγνώριση. Σημαντική είναι η σύγκριση με το Σχήμα 6.3, στο οποίο η εφαρμογή των τεχνικών προσαρμογής έχει πραγματοποιηθεί ανά ομιλητή, αλλά για τις ίδιες δύο εκφωνήσεις. Παρατηρείται πόσο διαφορετική είναι η μετακίνηση των τιμών στην περίπτωση της θυμωμένης εκφώνησης και συνεπώς η σημασία της διαφορετικής αντιμετώπισης κάθε συναισθηματικής κλάσης.

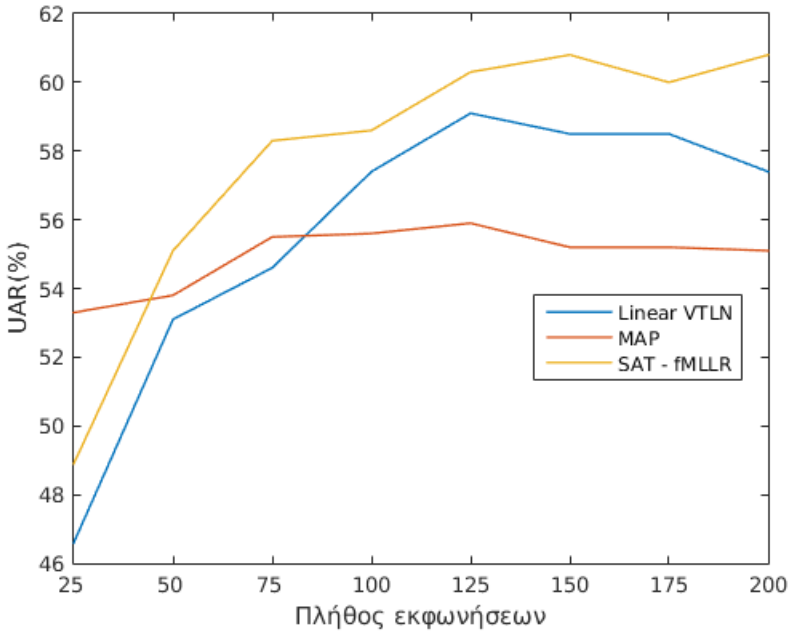
6.6 Προσαρμογή με Επίβλεψη

Μέχρι τώρα, οι τεχνικές Προσαρμογής του Ομιλητή εφαρμόστηκαν χωρίς επίβλεψη. Αξίζει να μελετηθεί η περίπτωση όπου διατίθεται μικρός αριθμός επισημειωμένων δεδομένων του ομιλητή αξιολόγησης. Βέβαια, αυτό μπορεί να μην είναι εφικτό σε πραγματικές εφαρμογές όπου ο ομιλητής είναι άγνωστος. Παρ' όλ' αυτά, η νέα τεχνολογία εξατομικευμένων συσκευών, όπως τα έξυπνα τηλέφωνα (smartphones), παρέχει τη δυνατότητα συλλογής προσωπικών φωνητικών δεδομένων του χρήστη [73]. Τα φωνητικά δεδομένα αυτά μπορούν να αξιοποιηθούν για την προσαρμογή ενός συστήματος στο συγκεκριμένο χρήστη, γνωρίζοντας την προσωπική του συναισθηματική έκφραση. Με αυτόν τον τρόπο, είναι εφικτή η βελτίωση προσωπικών φωνητικών εφαρμογών.

Για το σκοπό αυτό, μετά την εκπαίδευση του μοντέλου με τα δεδομένα των 9 ομιλητών, δοκιμάστηκε η προσαρμογή του με μικρό αριθμό επισημειωμένων εκφωνήσεων του δέκατου ομιλητή, και στη συνέχεια αξιολόγηση του προσαρμοσμένου μοντέλου με χρήση των υπόλοιπων εκφωνήσεων του δέκατου ομιλητή. Οι ετικέτες των εκφωνήσεων, που χρησιμοποιούνται για την αξιολόγηση του μοντέλου, θεωρούνται άγνωστες. Με τον τρόπο εκπαίδευσης που περιγράφτηκε, το μοντέλο προσαρμόζεται, έστω και σε μικρό βαθμό, στο νέο ομιλητή, ο οποίος θα χαρακτηρίζεται από τη δική του χροιά και προσωπική έκφραση, καθώς και στις νέες συνθήκες ηχογράφησης. Προφανώς, όσο περισσότερα επισημειωμένα δεδομένα δίνονται για προσαρμογή του μοντέλου, τόσο καλύτερη θα είναι και η μετέπειτα αναγνώριση συναισθήματος. Έτσι, ένα ερώτημα που προκύπτει αφορά τον αριθμό των επισημειωμένων δεδομένων που καθιστούν αποτελεσματική την προσαρμογή του συστήματος.

Στο Σχήμα 6.6 απεικονίζεται η μετρική UAR (%) ως προς τον αριθμό των επισημειωμένων εκφωνήσεων, οι οποίες δίνονται για προσαρμογή του μοντέλου GMM. Για τη προσαρμογή αυτή, εφαρμόζεται μία από τις 3 τεχνικές: Linear VTLN, MAP και SAT - fMLLR. Το υψηλότερο ποσοστό επιτυχίας παρατηρείται περίπου στις 125 εκφωνήσεις και για τις τρεις περιπτώσεις. Η τεχνική SAT δίνει για άλλη μια φορά τα καλύτερα αποτελέσματα, καθώς ήδη από τις 75 εκφωνήσεις για προσαρμογή, ξεπερνά το ποσοστό επιτυχίας του απλού μη-προσαρμοσμένου GMM μοντέλου και φτάνει το UAR 58.3%. Επίσης, με την αύξηση του αριθμού των εκφωνήσεων, επιτυγχάνει UAR 60.8%. Ωστόσο, και η Linear VTLN παρουσιάζει σημαντική αύξηση του

ποσοστού επιτυχίας, φτάνοντας το UAR 59.1% στις 125 εκφωνήσεις. Την πιο μικρή αύξηση παρουσιάζει η τεχνική MAP, η οποία φαίνεται να χρειάζεται ακόμα περισσότερες εκφωνήσεις για την αποτελεσματική προσαρμογή του μοντέλου στο νέο ομιλητή.

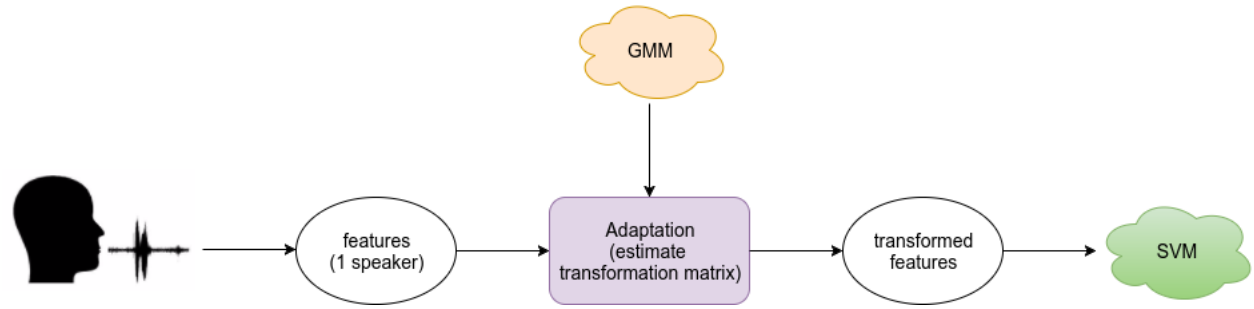


Σχήμα 6.6: *Μετρική UAR (%) ως προς τον αριθμό των εκφωνήσεων του ομιλητή αξιολόγησης, οι οποίες δίνονται για προσαρμογή του μοντέλου με μία από τις 3 τεχνικές: Linear VTLN, MAP, SAT - fMLLR.*

6.7 Ταξινόμηση με SVM

Κάποιες από τις τεχνικές Προσαρμογής του Ομιλητή που εφαρμόστηκαν στις προηγούμενες παραγράφους, εκτιμούν ουσιαστικά έναν αφινικό μετασχηματισμό των ακουστικών χαρακτηριστικών του ομιλητή, με σκοπό την προσαρμογή του στο μοντέλο. Τέτοιες είναι η fMLLR και η Linear VTLN. Ο αφινικός μετασχηματισμός που εκτιμάται επιδιώκει την ελαχιστοποίηση των διαφορών των εξαγόμενων χαρακτηριστικών μεταξύ των διαφορετικών ομιλητών. Η παρατήρηση αυτή οδηγεί στην ιδέα εφαρμογής του μετασχηματισμού αυτού στα αντίστοιχα χαρακτηριστικά, και στη μετέπειτα ταξινόμηση των μετασχηματισμένων χαρακτηριστικών με χρήση μοντέλου, διαφορετικού του GMM. Στο παρόν κεφάλαιο, χρησιμοποιείται SVM μοντέλο για την ταξινόμηση αυτή.

Η διαδικασία που ακολουθείται αναλύεται παρακάτω (βλ. Σχήμα 6.7). Αρχικά, εξάγονται τα ίδια ακουστικά χαρακτηριστικά, όπως στο μοντέλο GMM (βλ. παρ. 6.2.2), δηλαδή η Ενέργεια και 12 MFCCs. Τα χαρακτηριστικά αυτά κανονικοποιούνται με εφαρμογή της CMN ανά ομιλητή. Επιπλέον, υπολογίζεται η πρώτη και η δεύτερη παράγωγός τους. Τα τελικά 39 χαρακτηριστικά ανά πλαίσιο δίνονται για εκπαίδευση ενός Μοντέλου Μείγματος 6 Gaussian Συνιστωσών. Για την εκπαίδευση χρησιμοποιούνται τα δεδομένα των 9 ομιλητών, ενώ για την αξιολόγηση του μοντέλου χρησιμοποιούνται τα δεδομένα του δέκατου ομιλητή, όπως και πριν.



Σχήμα 6.7: *Ταξινόμηση με SVM των μετασχηματισμένων χαρακτηριστικών, μετά την εφαρμογή τεχνικής Προσαρμογής του Ομιλητή.*

Κατόπιν, εκτιμάται ένας αφινικός μετασχηματισμός των χαρακτηριστικών για κάθε ομιλητή, με σκοπό την προσαρμογή τους στο μοντέλο GMM. Η εκτίμηση του μετασχηματισμού αυτού, στην περίπτωση των δεδομένων εκπαίδευσης, μπορεί να γίνει είτε με επίβλεψη (SAT με fMLLR), είτε χωρίς επίβλεψη (απλή fMLLR). Στην περίπτωση των δεδομένων αξιολόγησης, οι ετικέτες θεωρούνται άγνωστες, οπότε προσαρμόζονται όπως και στην παράγραφο 6.4. Για τα μετασχηματισμένα χαρακτηριστικά των πλαισίων κάθε εκφώνησης υπολογίζονται τα 5 στατιστικά: ελάχιστη τιμή, μέγιστη τιμή, μέση τιμή, διάμεσος και τυπική απόκλιση. Με τον τρόπο αυτό, προκύπτουν τα χαρακτηριστικά για κάθε εκφώνηση, τα οποία δίνονται για εκπαίδευση ή αξιολόγηση ενός SVM μοντέλου με γραμμική συνάρτηση πυρήνα (linear kernel), ανάλογα με την περίπτωση. Για την υλοποίηση του SVM χρησιμοποιήθηκε το εργαλείο WEKA [40] και συγκεκριμένα ο αλγόριθμος ταξινόμησης SMO (Sequential Minimal Optimization).

	WAR (%)	UAR (%)
SVM	55.9	57.4
Linear VTLN + SVM	56.5	58.1
fMLLR + SVM	**56.9**	58.5
SAT - fMLLR + SVM	56.4	**58.8**

Πίνακας 6.4: *Μετρικές WAR (%) και UAR (%) για κάθε τεχνική Προσαρμογής του Ομιλητή, όταν πραγματοποιείται SVM ταξινόμηση των μετασχηματισμένων χαρακτηριστικών.*

Στον παραπάνω πίνακα (Πίνακας 6.4), παρουσιάζονται τα αποτελέσματα ταξινόμησης με SVM των μετασχηματισμένων χαρακτηριστικών, όταν εφαρμόζεται Linear VTLN, απλή fMLLR και SAT με fMLLR. Η πρώτη γραμμή αντιστοιχεί στο αποτέλεσμα της SVM ταξινόμησης χωρίς μετασχηματισμό των ακουστικών χαρακτηριστικών (μόνο εφαρμογή CMN, η οποία εφαρμόζεται σε όλες τις περιπτώσεις). Παρατηρείται ότι το SVM μοντέλο λειτουργεί καλύτερα από το GMM στο συγκεκριμένο πρόβλημα ταξινόμησης. Ακόμα και χωρίς μετασχηματισμό των χαρακτηριστικών, το SVM δίνει UAR 57.4%, ενώ το GMM δίνει UAR 55.2%. Επίσης, πάλι η τεχνική SAT με fMLLR οδηγεί στο πιο εύρωστο σύστημα (UAR 58.8%), με βελτίωση της αντίστοιχης μετρικής UAR κατά 2% περίπου, συγκριτικά με το GMM.

6.8 Συμπεράσματα

Από την ανάλυση των παραπάνω παραγράφων, συμπεραίνεται η χρησιμότητα των τεχνικών Προσαρμογής του Ομιλητή. Όσον αφορά την Αναγνώριση Συναισθήματος, οι συγκεκριμένες τεχνικές συνεισφέρουν στην προσαρμογή του συστήματος σε έναν καινούργιο ομιλητή ή σε διαφορετικές συνθήκες ηχογράφησης, έστω και σε μικρό βαθμό. Η σημαντικότερη βελτίωση παρατηρείται κατά την εφαρμογή προσαρμοζόμενης εκπαίδευσης SAT. Η τεχνική αυτή στοχεύει στο διαχωρισμό των φωνητικών ιδιαιτεροτήτων των ομιλητών από τα καθαρά φωνητικά χαρακτηριστικά, τα οποία θα οδηγήσουν στη σωστή αναγνώριση. Αποδεικνύεται ότι η προσέγγισή της λειτουργεί ιδιαίτερα καλά, τόσο κατά την αξιολόγηση με μοντέλο GMM όσο και με μοντέλο SVM, με χρήση των πινάκων μετασχηματισμού fMLLR. Ακόμα, δίνει σημαντική βελτίωση και κατά την προσαρμογή του μοντέλου GMM με επίβλεψη, με χρήση μικρού σχετικά αριθμού επισημειωμένων δεδομένων.

Σημαντική παρατήρηση, επίσης, αποτελεί η αξία της διαφορετικής αντιμετώπισης κάθε συναισθηματικής κλάσης. Ενώ είναι γενικά προφανές ότι κάθε ομιλητής έχει τη δική του προσωπική και συναισθηματική έκφραση, η διαφορά που παρατηρήθηκε κατά την προσαρμογή του μοντέλου ανά συναισθηματική κλάση θεωρείται αξιοσημείωτη. Δίνει το κίνητρο ανάπτυξης τρόπων εφαρμογής των τεχνικών προσαρμογής με βάση το συναίσθημα κάθε ομιλητή.

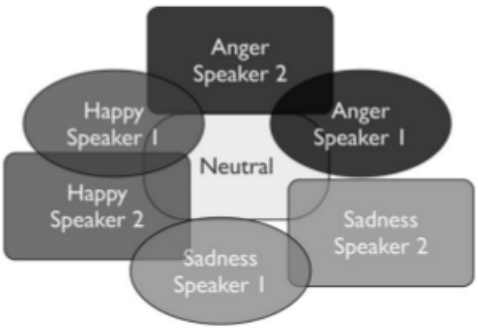
Τέλος, ιδιαίτερο ενδιαφέρον παρουσιάζει η χρήση των προσαρμοσμένων χαρακτηριστικών σε μοντέλο SVM. Εκτός ότι το SVM δείχνει καλύτερη απόδοση στο συγκεκριμένο πρόβλημα αναγνώρισης, παρατηρήθηκε σημαντική βελτίωση του ποσοστού επιτυχίας κατά την εφαρμογή των μετασχηματισμών fMLLR στα ακουστικά χαρακτηριστικά και στη συνέχεια χρήση τους για εκπαίδευση και αξιολόγηση του SVM. Προτείνεται, λοιπόν, η περαιτέρω δοκιμή και άλλων χαρακτηριστικών σε αυτό το μοντέλο. Καθώς η fMLLR είναι γενική ως τεχνική, είναι δυνατή η χρήση πλήθους χαρακτηριστικών εκτός τα MFCCs. Στο παρόν σύγγραμμα, έγινε αυτή η επιλογή χαρακτηριστικών με σκοπό την εύκολη σύγκριση των μεθόδων προσαρμογής.

Κεφάλαιο 7

Επαναληπτική Προσαρμογή των Χαρακτηριστικών

7.1 Εισαγωγή

Μια διαφορετική προσέγγιση για την προσαρμογή ενός ομιλητή σε ένα σύστημα αναγνώρισης συναισθήματος προτείνεται από το [49]. Σκοπός είναι ο διαχωρισμός των συναισθηματικών εκφωνήσεων από τις ουδέτερες. Η διαδικασία που ακολουθείται επιδιώκει την ελαχιστοποίηση των διαφορών της ουδέτερης ομιλίας μεταξύ των ομιλητών, με ταυτόχρονη όμως διατήρηση της διακριτότητας των συναισθηματικών κλάσεων. Η ιδέα αυτή απεικονίζεται στο Σχήμα 7.1, όπου μετά την κανονικοποίηση των χαρακτηριστικών που εφαρμόζεται, έχει επιτευχθεί ταύτιση των ουδέτερων χαρακτηριστικών μεταξύ των διαφορετικών ομιλητών και διατήρηση των διαφορών των συναισθηματικών κλάσεων. Αποτέλεσμα είναι μόνο η μείωση της μεταβλητότητας των χαρακτηριστικών που οφείλεται στη διαφορετικότητα των ομιλητών ή των συνθηκών ηχογράφησης, και όχι αυτής που προκαλεί η έκφραση συναισθήματος. Με αυτόν τον τρόπο, είναι εφικτή η διάκριση οποιασδήποτε συναισθηματικής εκφώνησης σε σύγκριση με τις ουδέτερες.



Σχήμα 7.1: *Συναισθηματικές κλάσεις στο χώρο των χαρακτηριστικών μετά την κανονικοποίηση, όπως αναφέρεται στο [49]. (Η γραφική απεικόνιση προέρχεται από το [49].)*

Στο παραπάνω σχήμα παρατηρείται για ακόμη μια φορά η μεταβλητότητα των συναισθημάτων μεταξύ διαφορετικών ατόμων, η οποία καθιστά ιδιαίτερα δύσκολη την αναγνώριση ενός συγκεκριμένου συναισθήματος σε έναν καινούργιο ομιλητή. Έτσι, σε πολλές εφαρμογές, ε-

ίναι αρχικά επιθυμητή η ανίχνευση της ουδέτερής του έκφρασης. Αντίστοιχα, χρήσιμη είναι η ανίχνευση της οποιαδήποτε μη ουδέτερης κατάστασης, με σκοπό τη βελτίωση της αλληλεπίδρασης ανθρώπου-μηχανής. Σε αυτήν την κατεύθυνση, το συγκεκριμένο άρθρο προτείνει μια επαναληπτική διαδικασία χωρίς επίβλεψη, που αποσκοπεί στην παραπάνω ιδέα ταύτισης της ουδέτερης ομιλίας. Η διαδικασία αυτή επιλέγει, αρχικά, τις ουδέτερες εκφωνήσεις ενός ομιλητή και στη συνέχεια, υπολογίζει τις παραμέτρους κανονικοποίησης με βάση μόνο τις εκφωνήσεις αυτές. Η κανονικοποίηση εφαρμόζεται στο σύνολο των δεδομένων, που αποτελούνται από την ουδέτερη και τη συναισθηματική κλάση. Επαναληπτικά, τα δεδομένα ταξινομούνται ξανά μέχρι τη σύγκλιση, και κανονικοποιούνται με βάση μόνο την ουδέτερη ομιλία. Υποστηρίζεται ότι, σε αντίθεση με την προτεινόμενη μέθοδο κανονικοποίησης, η όποια καθολική κανονικοποίηση των χαρακτηριστικών είναι δυνατόν να επηρεάσει τη διακριτότητα των συναισθηματικών κλάσεων.

Στο συγκεκριμένο άρθρο, ως τεχνική κανονικοποίησης των χαρακτηριστικών χρησιμοποιήθηκε η Z-Normalization και SVM ως μοντέλο για την αναγνώριση. Δεδομένου ότι ο μετασχηματισμός των χαρακτηριστικών είναι αφινικός, θα διατηρηθεί η διαφορετικότητα των συναισθηματικών κλάσεων, μετά την επιβολή του. Με βάση τη συγκεκριμένη αντιμετώπιση, εδώ, αναπτύσσεται ένα παρόμοιο σύστημα, με χρήση, όμως, της τεχνικής προσαρμογής fMLLR για την εκτίμηση του αφινικού μετασχηματισμού για κάθε ομιλητή. Τόσο η εκτίμηση του μετασχηματισμού των χαρακτηριστικών, όσο και η ταξινόμηση πραγματοποιούνται με βάση Μοντέλο Μείγματος Gaussian Συνιστωσών (GMM). Στις επόμενες παραγράφους, παρουσιάζονται οι λεπτομέρειες του συστήματος και τα αποτελέσματά του. Ως εργαλείο για την εξαγωγή των ακουστικών χαρακτηριστικών και την υλοποίηση του GMM και των τεχνικών προσαρμογής, χρησιμοποιήθηκε το Kaldi, όπως και στο προηγούμενο κεφάλαιο.

7.2 Περιγραφή Συστήματος

7.2.1 Δεδομένα

Ως βάση δεδομένων χρησιμοποιήθηκε το IEMOCAP, όπως και στο προηγούμενο κεφάλαιο (βλ. παρ. 6.2.1), και μάλιστα το ίδιο σύνολο των εκφωνήσεων. Στο παρόν κεφάλαιο, οι 5531 εκφωνήσεις χωρίστηκαν σε δύο κλάσεις:

- *συναισθηματική*: χαρά (595 εκφωνήσεις), ενθουσιασμός (1041 εκφωνήσεις), λύπη (1084 εκφωνήσεις), θυμός (1103 εκφωνήσεις)
- *ουδέτερη*: ουδέτερο (1708 εκφωνήσεις)

Παρατηρείται ότι η συναισθηματική κλάση περιλαμβάνει περίπου 3 φορές περισσότερες εκφωνήσεις, συγκριτικά με την ουδέτερη. Σε αντίθεση με το [49], δεν επιλέχθηκαν δεδομένα, με σκοπό τον ίσο αριθμό εκφωνήσεων στις δύο κλάσεις. Επίσης, δεν αποκλείστηκαν εκφωνήσεις, στις οποίες έχει αποδοθεί έστω μία ουδέτερη ετικέτα από κάποιον επισημειωτή, αλλά έχουν καταταχθεί σε συναισθηματική κλάση με βάση την πλειοψηφία. Οι παραπάνω υποθέσεις έγιναν, έτσι ώστε το σύστημα να αξιολογηθεί σε πραγματικές καταστάσεις, όπου μπορεί να προκύψουν λάθη και μη ισορροπία των κλάσεων.

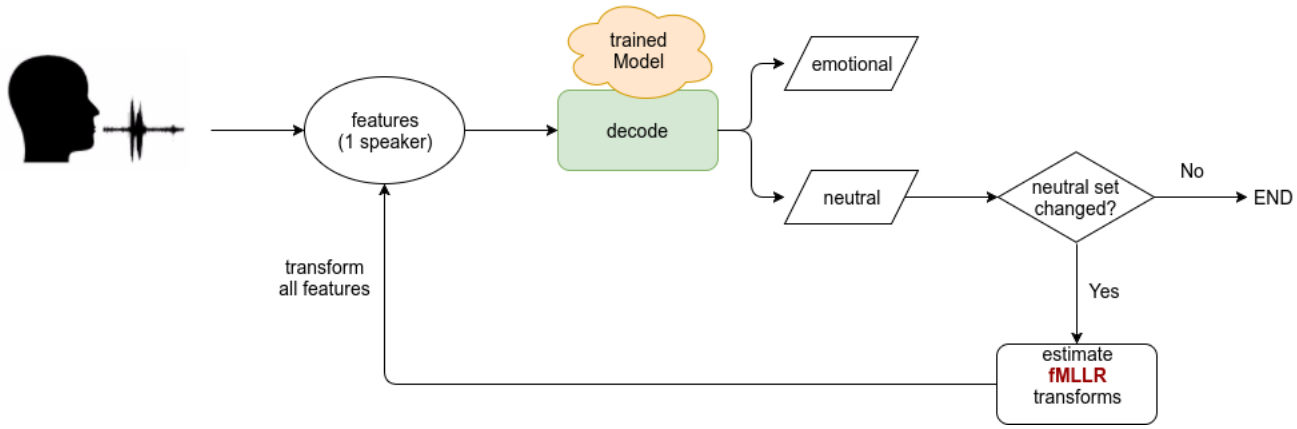
Όσον αφορά το χωρισμό των δεδομένων, με σκοπό την αξιολόγηση του συστήματος, ακολουθήθηκε *10-fold leave-one-out* προσέγγιση. Κυκλικά, δηλαδή, χρησιμοποιούνται οι εκφωνήσεις των 9 ομιλητών για την εκπαίδευση ενός μοντέλου και πραγματοποιείται αξιολόγηση με

χρήση των δεδομένων του δέκατου ομιλητή. Οι μετρικές που παρατίθενται, προκύπτουν ως η μέση τιμή των αντίστοιχων αποτελεσμάτων των 10 αυτών πειραμάτων.

7.2.2 Ακουστικά Χαρακτηριστικά

Ως αρχικά χαρακτηριστικά θεωρούνται η Ενέργεια και οι MFCC συνιστώσες 1-12, όπως και στο προηγούμενο κεφάλαιο (βλ. παρ. 6.2.2). Επιπλέον, εφαρμόζεται CMN ανά ομιλητή και υπολογίζονται η πρώτη και η δεύτερη παράγωγός τους. Συνολικά, προκύπτουν 39 ακουστικά χαρακτηριστικά ανά πλαίσιο.

7.2.3 Περιγραφή Διαδικασίας



Σχήμα 7.2: *Επαναληπτική Προσαρμογή των Χαρακτηριστικών: Επαναληπτικός υπολογισμός fMLLR μετασχηματισμών με βάση τις ουδέτερες εκφωνήσεις, οι οποίες προβλέπονται σε κάθε επανάληψη, και εφαρμογή των μετασχηματισμών αυτών σε όλα τα χαρακτηριστικά.*

Η διαδικασία που ακολουθείται, εφαρμόζει ουσιαστικά Επαναληπτική Προσαρμογή των Χαρακτηριστικών και απεικονίζεται στο Σχήμα 7.2. Αρχικά, εξάγονται τα χαρακτηριστικά για κάθε εκφώνηση κάθε ομιλητή. Όπως αναφέρθηκε και παραπάνω, οι εκφωνήσεις των 9 ομιλητών χρησιμοποιούνται για την εκπαίδευση ενός μοντέλου και η αξιολόγηση πραγματοποιείται στον δέκατο ομιλητή. Με βάση την πρόβλεψη που προκύπτει για τις ετικέτες των εκφωνήσεων του συγκεκριμένου ομιλητή αξιολόγησης (ουδέτερη ή συναισθηματική), δημιουργείται μια ομάδα που περιλαμβάνει αποκλειστικά τις ουδέτερες από αυτές. Αυτή η ομάδα χρησιμοποιείται για την εκτίμηση των fMLLR μετασχηματισμών, με βάση το εκπαιδευμένο μοντέλο. Οι μετασχηματισμοί αυτοί εφαρμόζονται στα χαρακτηριστικά των δεδομένων του ομιλητή, και των 2 κλάσεων (συναισθηματική και ουδέτερη). Τα κανονικοποιημένα χαρακτηριστικά χρησιμοποιούνται για την αναγνώριση της επόμενης επανάληψης, με βάση πάλι το αρχικό εκπαιδευμένο μοντέλο. Η διαδικασία επαναλαμβάνεται μέχρι την επαλήθευση κάποιου κριτηρίου τερματισμού. Στο Σχήμα 7.2, ως κριτήριο έχει τεθεί η αλλαγή ή όχι της ουδέτερης ομάδας (neutral set), δηλαδή της πρόβλεψης των συναισθηματικών κλάσεων, σε σχέση με την προηγούμενη επανάληψη. Δυνατός είναι και ο ορισμός ενός μέγιστου αριθμού επαναλήψεων.

7.2.4 Μοντέλο Μείγματος Gaussian Συνιστωσών

Όσον αφορά την εκπαίδευση του μοντέλου με τα δεδομένα των 9 ομιλητών (*trained Model* στο Σχήμα 7.2), επιχειρήθηκαν 2 περιπτώσεις:

- *GMM - fMLLR από ουδέτερη ομάδα*
- *GMM - SAT (fMLLR)*

Και στις δύο περιπτώσεις υλοποιήθηκε Μοντέλο Μείγματος 13 Gaussian Συνιστωσών για κάθε κλάση. Η διαφορά εντοπίζεται στον τρόπο εκτίμησης του αφινικού μετασχηματισμού fMLLR, ο οποίος εφαρμόζεται στα ακουστικά χαρακτηριστικά με σκοπό την εκπαίδευση του GMM μοντέλου. Στην πρώτη περίπτωση, ο fMLLR πίνακας μετασχηματισμού εκτιμάται με βάση τις ουδέτερες εκφωνήσεις κάθε ομιλητή, παρόμοια δηλαδή με το [49]. Εφόσον, για τους 9 ομιλητές εκπαίδευσης, οι ετικέτες είναι διαθέσιμες, είναι γνωστή η ουδέτερη ομάδα εκφωνήσεων για κάθε έναν από αυτούς. Έτσι, υπολογίζεται με ακρίβεια ο μετασχηματισμός κάθε ομιλητή με βάση αυτήν την ομάδα και στη συνέχεια, εφαρμόζεται σε όλα τα χαρακτηριστικά του, για την εκπαίδευση του GMM. Για την εκτίμηση των fMLLR μετασχηματισμών, έχει αρχικά εκπαιδευτεί ένα GMM με χρήση των αρχικών χαρακτηριστικών. Στη δεύτερη περίπτωση, μετά από εκπαίδευση ενός αρχικού GMM με χρήση των αρχικών χαρακτηριστικών, εφαρμόζεται η τεχνική προσαρμοζόμενης εκπαίδευσης ανά ομιλητή SAT (Speaker Adaptive Training). Για κάθε ομιλητή, εκτιμάται επαναληπτικά ένας fMLLR μετασχηματισμός για τα χαρακτηριστικά του συνόλου των εκφωνήσεών του. Σκοπός είναι η μεγιστοποίηση της πιθανοφάνειας των δεδομένων του. Σε αυτή την περίπτωση, δεν αξιοποιείται η ιδέα του υπολογισμού των παραμέτρων κανονικοποίησης με βάση μόνο την ουδέτερη ομιλία, όπως στην πρώτη περίπτωση. Όμως, δοκιμάζεται η επίδοση της τεχνικής SAT και σε αυτό το σύστημα, καθώς έδωσε σημαντική βελτίωση στο προηγούμενο κεφάλαιο.

Όσον αφορά τον τρόπο αναγνώρισης των ετικετών του ομιλητή αξιολόγησης (*decode* στο Σχήμα 7.2), πραγματοποιήθηκε απλή αξιολόγηση του GMM μοντέλου που προέκυψε κατά την εκπαίδευση, είτε με την πρώτη είτε με τη δεύτερη μέθοδο. Τα χαρακτηριστικά του ομιλητή αξιολόγησης κανονικοποιούνται με εφαρμογή του υπολογισμένου fMLLR μετασχηματισμού, με βάση την ουδέτερη ομάδα της προηγούμενης επανάληψης. Για την πρώτη επανάληψη, απαιτείται κάποια αρχικοποίηση των παραμέτρων του μετασχηματισμού αυτού. Επιλέχθηκε η εκτίμηση πίνακα fMLLR με βάση τα συνολικές εκφωνήσεις του ομιλητή, δηλαδή κλασσική τεχνική fMLLR (καθολική ανά ομιλητή).

7.3 Σύγκριση με GMM

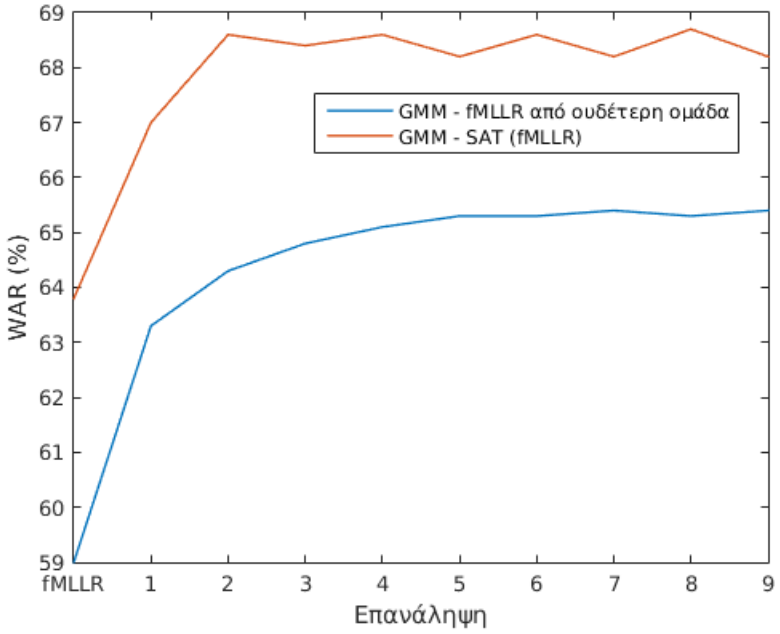
Αρχικά, για να υπάρχει μία σύγκριση των αποτελεσμάτων της επαναληπτικής μεθόδου, αναγράφονται οι μετρικές WAR και UAR, που προκύπτουν κατά την αξιολόγηση απλού Μοντέλου Μείγματος 13 Gaussian Συνιστωσών, για τις δύο κλάσεις (συναισθηματική και ουδέτερη), εκπαιδευμένου με τα αρχικά χαρακτηριστικά (βλ. παρ. 7.2.2). Εν συνεχεία, εφαρμόζεται η τεχνική fMLLR με τον κλασσικό τρόπο, καθολική ανά ομιλητή αξιολόγησης. Επίσης, παρατίθενται τα αντίστοιχα αποτελέσματα, αν το σύστημα έχει εκπαιδευτεί με χρήση της τεχνικής προσαρμοζόμενης εκπαίδευσης SAT με fMLLR ανά ομιλητή.

	WAR (%)	UAR (%)
CMN	65.4	65.7
CMN + fMLLR	61.4	65.3
CMN + SAT - fMLLR	63.8	66.0

Πίνακας 7.1: *Μετρικές WAR (%) και UAR (%) που προκύπτουν από GMM για 2 κλάσεις (συναισθηματική και ουδέτερη), για 3 τεχνικές Προσαρμογής του Ομιλητή.*

7.4 Αποτελέσματα

Στο Σχήμα 7.3, απεικονίζεται το ποσοστό επιτυχίας που προκύπτει από την αξιολόγηση του συστήματος Επαναληπτικής Προσαρμογής των Χαρακτηριστικών (βλ. παρ. 7.2) ως προς τον αριθμό των επαναλήψεων, και για τις δύο περιπτώσεις εκπαίδευσης του μοντέλου GMM. Για την απεικόνιση αυτή ως κριτήριο τερματισμού θεωρήθηκαν οι 10 επαναλήψεις. Η πρώτη τιμή του διαγράμματος αντιστοιχεί στην επανάληψη *0*, όπου η αναγνώριση γίνεται με αρχικοποίηση των παραμέτρων μετασχηματισμού. Η αρχικοποίηση επιλέχθηκε να γίνεται με εκτίμηση καθολικού fMLLR μετασχηματισμού των χαρακτηριστικών του ομιλητή αξιολόγησης με βάση το εκπαιδευμένο GMM, όπως αναφέρθηκε και παραπάνω.

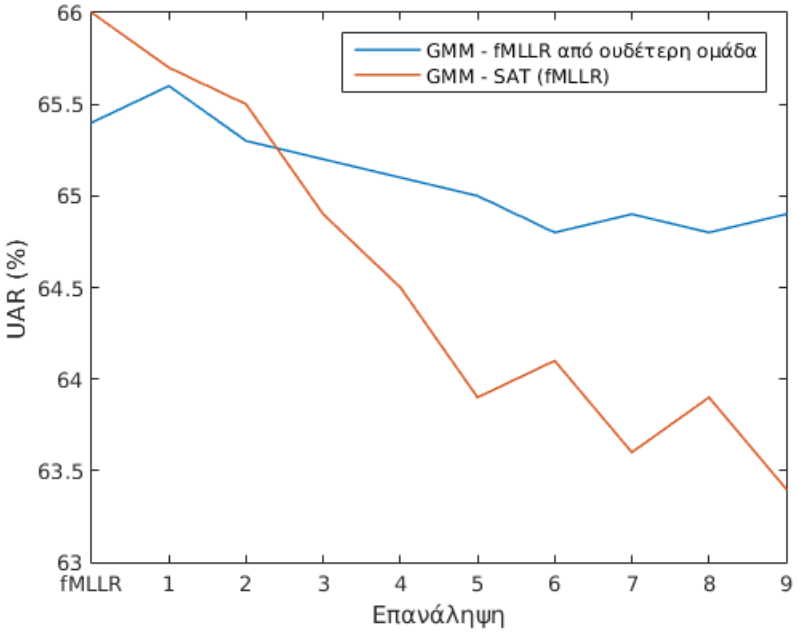


Σχήμα 7.3: *Μετρική WAR (%) ως προς τον αριθμό των επαναλήψεων, κατά την αξιολόγηση του συστήματος Επαναληπτικής Προσαρμογής των Χαρακτηριστικών, για τις 2 περιπτώσεις εκπαίδευσης του μοντέλου GMM.*

Παρατηρείται σημαντική αύξηση της μετρικής WAR, κυρίως μέχρι την τρίτη περίπου επανάληψη και για τις δύο περιπτώσεις. Σημαντικά καλύτερα αποτελέσματα παρουσιάζει η τεχνική SAT, με χρήση της οποίας πετυχαίνεται WAR 68.7% στην όγδοη επανάληψη, δίνοντας βελτίω-

ση 3% περίπου από το απλό GMM και αύξηση κατά 5% περίπου σε σχέση με την κλασσική αναγνώριση συστήματος SAT με fMLLR. Όσον αφορά τη σύγκλιση των δύο περιπτώσεων, καλύτερη φαίνεται να παρουσιάζει η τεχνική εφαρμογής fMLLR με βάση μόνο την ουδέτερη ομάδα. Ωστόσο, η αστάθεια που φαίνεται να έχει η τεχνική SAT είναι της τάξης του 0.4%, το οποίο αντιστοιχεί σε μόλις 2 δείγματα στα συνολικά 500-600 περίπου ανά ομιλητή, και άρα δεν είναι στατιστικά σημαντική.

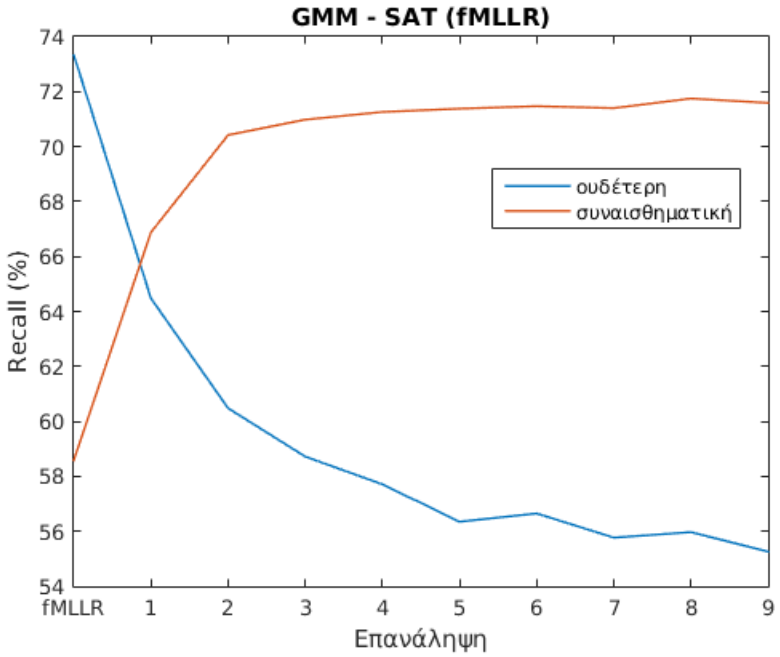
Στο Σχήμα 7.4, απεικονίζεται η μετρική UAR ως προς τον αριθμό των επαναλήψεων, για την ίδια περίπτωση πειράματος. Παρατηρείται ελαφρά πτώση της συγκεκριμένης μετρικής, κυρίως όταν εφαρμόζεται η τεχνική εκπαίδευσης SAT. Η πτώση αυτή οφείλεται στη μη ισορροπία των δύο κλάσεων, όπου η συναισθηματική κλάση περιλαμβάνει περίπου τρεις φορές περισσότερα δεδομένα από την ουδέτερη, όπως αναφέρθηκε και παραπάνω. Βέβαια, η πτώση αυτή δεν κρίνεται τόσο σημαντική, συγκριτικά με την αντίστοιχη αύξηση της μετρικής WAR. Για παράδειγμα, στην τεχνική SAT, ήδη στη δεύτερη επανάληψη, έχει επιτευχθεί WAR 68.6% (αύξηση κατά περίπου 5%) και UAR 65.5% (μείωση κατά 0.5%). Γενικά, συμπεραίνεται ότι η πραγματοποίηση 2-3 επαναλήψεων της συγκεκριμένης μεθόδου οδηγεί σε σημαντική βελτίωση της απόδοσης του συστήματος.



Σχήμα 7.4: *Μετρική UAR (%) ως προς τον αριθμό των επαναλήψεων, κατά την αξιολόγηση του συστήματος Επαναληπτικής Προσαρμογής των Χαρακτηριστικών, για τις 2 περιπτώσεις εκπαίδευσης του μοντέλου GMM.*

Για την περαιτέρω αξιολόγηση της μετρικής UAR, παρατίθεται το διάγραμμα της ανάκλησης για κάθε κλάση, ως προς τον αριθμό των επαναλήψεων, όταν εφαρμόζεται τεχνική SAT για την εκπαίδευση του μοντέλου GMM (βλ. Σχήμα 7.5). Με τον τρόπο αυτό, φαίνεται η βασική αιτία της μείωσης της μετρικής UAR, εφόσον αυτή συνιστά τη μέση τιμή των δύο ανακλήσεων που απεικονίζονται. Παρατηρείται ότι κατά την αύξηση του αριθμού των επαναλήψεων, σημειώνεται μείωση της ανάκλησης της ουδέτερης κλάσης και αύξηση της ανάκλησης της συναισθηματικής

κλάσης. Συμπερασματικά, έχει επιτευχθεί καλύτερη ανίχνευση των συναισθηματικών εκφωνήσεων, εις βάρος των ουδέτερων. Το αποτέλεσμα αυτό είναι πιθανότατα επιθυμητό για τη βελτίωση εφαρμογών αλληλεπίδρασης ανθρώπου-μηχανής.



Σχήμα 7.5: *Γραφική απεικόνιση της ανάκλησης (%) (recall) για τις 2 κλάσεις (συναισθηματική και ουδέτερη), ως προς τον αριθμό των επαναλήψεων, κατά την αξιολόγηση του συστήματος Επαναληπτικής Προσαρμογής των Χαρακτηριστικών, όταν εφαρμόζεται τεχνική SAT για την εκπαίδευση του μοντέλου GMM.*

7.4.1 Κριτήριο Τερματισμού

Μια σημαντική παρατήρηση κατά την ανάλυση των αποτελεσμάτων είναι η διαφορετική απόδοση του κάθε ομιλητή. Κάθε ομιλητής έχει μέγιστο ποσοστό επιτυχίας σε διαφορετική επανάληψη, με αποτέλεσμα ο μέσος όρος να μη δείχνει την πραγματική εικόνα για κάθε ξεχωριστό πείραμα. Με σκοπό την περαιτέρω βελτίωση της εικόνας, δοκιμάστηκε ο τερματισμός των επαναλήψεων στο καλύτερο δυνατό αποτέλεσμα για κάθε ομιλητή, με εφαρμογή κριτηρίου τερματισμού. Σε κάθε επανάληψη i, συγκρίνονται οι εκφωνήσεις που προβλέφθηκαν ως ουδέτερες (ουδέτερη ομάδα), σε σχέση με τις αντίστοιχες της προηγούμενης επανάληψης $i-1$. Αν κάποια από αυτές, δεν υπήρχε στην ουδέτερη ομάδα της προηγούμενης επανάληψης, ένας μετρητής διαφοράς d_i αυξάνεται κατά 1. Έχοντας συγκρίνει τις ουδέτερες ομάδες, ορίστηκε το κριτήριο τερματισμού:

$$d_i \leq 10 \quad \&\& \quad (d_i > d_{i-1} \quad \text{ή} \quad d_i = 0) \tag{7.4.1}$$

Δηλαδή, η επαναληπτική διαδικασία τερματίζεται όταν η διαφορά της ουδέτερης ομάδας σε σχέση με αυτήν της προηγούμενης επανάληψης έχει μειωθεί κάτω από ένα κατώφλι (10) και σε περίπτωση που η διαφορά αυτή αρχίσει να αυξάνεται ή γίνει ίση με 0. Το τελικό αποτέλεσμα είναι η κατηγοριοποίηση που μόλις προέκυψε. Ως κατώφλι θα μπορούσε να οριστεί και κάποια

άλλη τιμή. Η τιμή αυτή δόθηκε, παρατηρώντας τις διαφορές που προκύπτουν στα πειράματα. Επίσης, θεωρώντας ότι είναι διαθέσιμες περίπου 500-600 εκφωνήσεις ανά ομιλητή, το ποσοστό των διαφορετικών ετικετών προκύπτει μικρότερο του 2%. Σε περίπτωση που η διαφορά d_i δεν μειωθεί κάτω από το κατώφλι 10, κατά τη διάρκεια των 10 επαναλήψεων, ως τελικό αποτέλεσμα θεωρείται αυτό όπου προέκυψε η μικρότερη d_i. Με βάση την παραπάνω περιγραφή, παρουσιάζονται παρακάτω τα αποτελέσματα για την περίπτωση εκπαίδευσης του μοντέλου με χρήση της τεχνικής SAT με fMLLR.

	WAR (%)	UAR (%)
GMM - SAT (fMLLR)	**69.8**	**65.9**

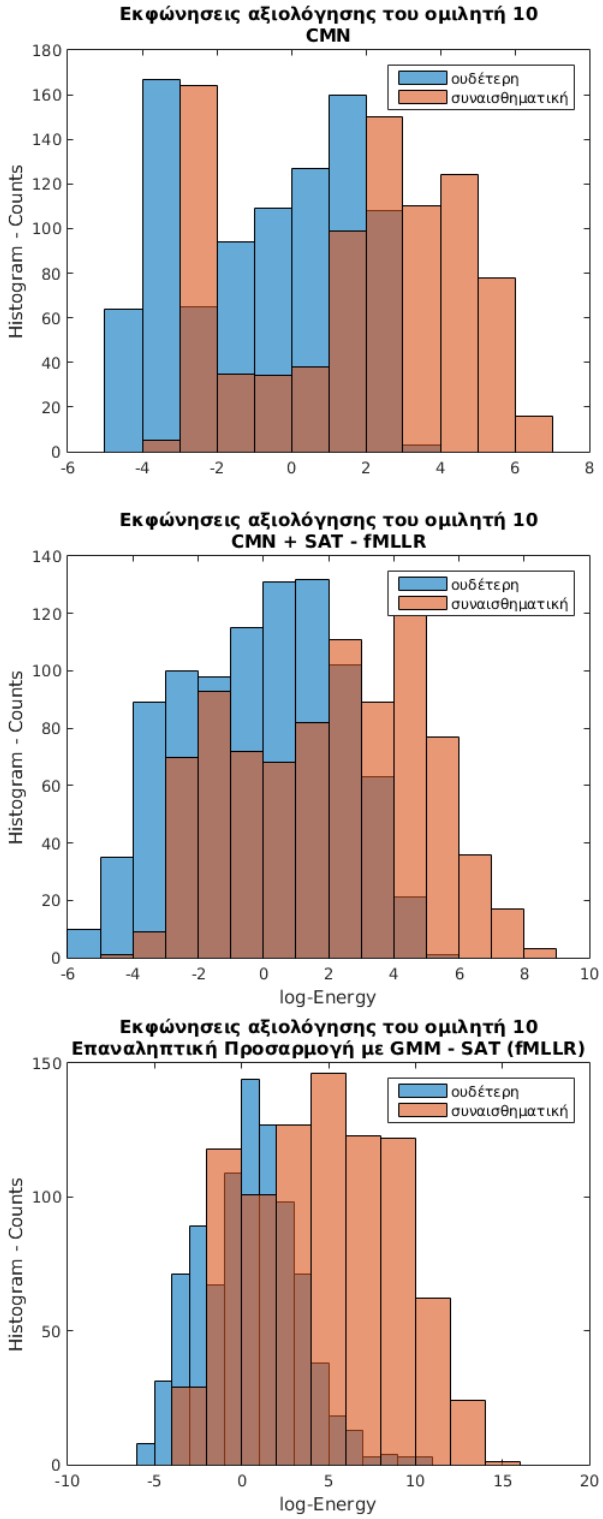
Πίνακας 7.2: *Μετρικές WAR (%) και UAR (%) για το σύστημα Επαναληπτικής Προσαρμογής των Χαρακτηριστικών, όταν εφαρμόζεται το κριτήριο τερματισμού της Σχέσης (7.4.1).*

Παρατηρείται εμφανής βελτίωση του ποσοστού επιτυχίας στην παραπάνω περίπτωση, όταν εφαρμόζεται το κριτήριο τερματισμού της Σχέσης (7.4.1). Η βελτίωση της μετρικής WAR ανέρχεται πάνω από 4% συγκριτικά με το απλό GMM και στο 6% σε σχέση με την κλασσική αναγνώριση με fMLLR. Επίσης, σημειώνεται μικρή βελτίωση και όσον αφορά τη μετρική UAR, δείχνοντας έτσι τη σημασία της εφαρμογής κριτηρίου τερματισμού.

7.4.2 Σύγκριση Ιστογραμμάτων

Παρακάτω, παρατίθενται προς σύγκριση τα ιστογράμματα της Ενέργειας για δύο εκφωνήσεις του δέκατου ομιλητή. Η μία εκφώνηση αντιστοιχεί στην ουδέτερη κλάση (μπλε) και η άλλη στην συναισθηματική κλάση (κόκκινο), και συγκεκριμένα έχει ετικέτα *θυμός*. Η πρώτη εικόνα από τις τρεις, απεικονίζει την Ενέργεια μετά την εφαρμογή CMN ανά ομιλητή και πριν οποιαδήποτε προσαρμογή των χαρακτηριστικών. Η δεύτερη εικόνα απεικονίζει την αντίστοιχη Ενέργεια μετά την εφαρμογή κλασσικής fMLLR τεχνικής, καθολικής ανά ομιλητή, με σκοπό την αξιολόγηση του συστήματος SAT. Τέλος, η τρίτη εικόνα αντιστοιχεί στην επαναληπτική μέθοδο προσαρμογής που αναπτύχθηκε στο παρόν κεφάλαιο. Συγκεκριμένα απεικονίζει την Ενέργεια μετά την εφαρμογή fMLLR μετασχηματισμού, υπολογισμένου με βάση τις ουδέτερες εκφωνήσεις, κατά την τρίτη επανάληψη, όπου επαληθεύεται το κριτήριο τερματισμού. Το μοντέλο GMM έχει εκπαιδευτεί με τεχνική SAT.

Παρατηρείται ότι η προτεινόμενη επαναληπτική μέθοδος επιτυγχάνει καλύτερο διαχωρισμό των δύο κλάσεων, καθώς μειώνει την επικάλυψη των δύο κατανομών, συγκρίνοντας με τις δύο πρώτες εικόνες. Η συναισθηματική κλάση έχει μετακινηθεί προς πιο υψηλές τιμές της Ενέργειας σε λογαριθμική κλίμακα, συγκριτικά με την ουδέτερη κλάση. Αποτέλεσμα είναι η καλύτερη μοντελοποίηση των δύο κατανομών και η βελτίωση του ποσοστού επιτυχίας, σε σχέση με την απλή αξιολόγηση του GMM μοντέλου που έχει εκπαιδευτεί με SAT.



Σχήμα 7.6: *Ιστογράμματα της Ενέργειας για 2 εκφωνήσεις (ουδέτερη και συναισθηματική) του ομιλητή 10: μετά από εφαρμογή CMN (πρώτο), μετά από εφαρμογή fMLLR (δεύτερο) και μετά την τελική εφαρμογή fMLLR, όταν τερματίζονται οι επαναλήψεις λόγω του κριτηρίου τερματισμού (τρίτο).*

7.5 Συμπεράσματα

Στο παρόν κεφάλαιο, αναπτύχθηκε μια επαναληπτική μέθοδος Προσαρμογής των Χαρακτηριστικών ενός ομιλητή. Ο κύριος στόχος της είναι η ανίχνευση των συναισθηματικών εκφωνήσεων και ο διαχωρισμός τους από τις ουδέτερες. Η ανίχνευση αυτή μπορεί να βελτιώσει τις δυνατότητες μιας πραγματικής εφαρμογής που περιλαμβάνει αλληλεπίδραση ανθρώπου-μηχανής, καθώς δίνει μια ένδειξη της συναισθηματικής φόρτισης ενός ατόμου. Πλεονεκτήματα που παρουσιάζει η συγκεκριμένη μέθοδος είναι η ικανότητα προσαρμογής του συστήματος στον καινούργιο ομιλητή, ανιχνεύοντας την ουδέτερη προσωπική του έκφραση, αλλά και η αντίληψη οποιασδήποτε πρόκλησης συναισθήματος. Σημαντικό στοιχείο συνιστά η εφαρμογή της χωρίς επίβλεψη, καθώς ανιχνεύει τις ουδέτερες εκφωνήσεις επαναληπτικά, βελτιώνοντας τον τελικό διαχωρισμό. Έτσι, μπορεί να εφαρμοστεί σε πραγματικές καταστάσεις, όπου κάθε καινούργιος ομιλητής είναι άγνωστος.

Σύμφωνα με τα αποτελέσματα της μεθόδου, μεγαλύτερο ποσοστό επιτυχίας εμφανίζεται όταν το εκπαιδευμένο μοντέλο προκύπτει με εφαρμογή της τεχνικής SAT (Speaker Adaptive Training). Η τεχνική εκπαίδευσης αυτή συγκρίνεται με την εφαρμογή μετασχηματισμού fMLLR που έχει εκτιμηθεί με βάση μόνο τις ουδέτερες εκφωνήσεις εκπαίδευσης κάθε ομιλητή, όπως υλοποιείται στο [49]. Τελικά, αποδεικνύεται ότι η συγκεκριμένη τεχνική προσαρμοζόμενης εκπαίδευσης (SAT), εμφανίζει πλεονεκτήματα για ακόμη μία φορά. Κατά την αξιολόγηση, αξιοποιείται η ιδέα της εκτίμησης του μετασχηματισμού των χαρακτηριστικών με βάση μόνο τις ουδέτερες εκφωνήσεις που προβλέπονται σε κάθε επανάληψη. Με βάση το κριτήριο τερματισμού που αναλύθηκε, προκύπτει τελική βελτίωση του ποσοστού επιτυχίας πάνω από 4% σε σχέση με το απλό GMM και 6% σε σχέση με την κλασσική αναγνώριση με fMLLR. Αν συγκριθεί το τελικό αποτέλεσμα με το καλύτερο του [49], προκύπτει βελτίωση σχεδόν 3%, αν και δεν είναι άμεσα συγκρίσιμα λόγω του διαφορετικού συνόλου εκφωνήσεων που χρησιμοποιούνται (βλ. παρ. 7.2.1).

Κεφάλαιο 8

Επίλογος

8.1 Σύνοψη Εργασίας και Συμπεράσματα

Στο παρόν κεφάλαιο, παρουσιάζονται συνοπτικά τα συμπεράσματα της συγκεκριμένης εργασίας, καθώς και πιθανές προεκτάσεις της. Ο βασικός της στόχος ήταν η μείωση όλων εκείνων των διαφορών των ομιλητών, οι οποίες επηρεάζουν τα εξαγόμενα ακουστικά χαρακτηριστικά και κατ'επέκταση την απόδοση ενός συστήματος αναγνώρισης συναισθήματος από φωνή. Οι διαφορετικοί ομιλητές διακρίνονται από βιολογικά και κοινωνικο-πολιτισμικά στοιχεία, τα οποία συχνά παίζουν σημαντικό ρόλο στην έκφραση συναισθημάτων. Επίσης, παρατηρείται μεταβολή των ακουστικών χαρακτηριστικών και λόγω των συνθηκών ηχογράφησης. Το αποτέλεσμα είναι η χαμηλή απόδοση ενός συστήματος, κατά την αξιολόγησή του με χρήση δεδομένων από διαφορετικό ομιλητή ή περιβάλλον ηχογράφησης, συγκριτικά με τα δεδομένα εκπαίδευσης.

Μια πρώτη προσέγγιση για τη μείωση των παραπάνω διαφορών είναι η κανονικοποίηση των χαρακτηριστικών (βλ. κεφ. 4), με χρήση απλών τεχνικών. Παραδείγματα τέτοιων τεχνικών είναι οι ακόλουθες: *Z-Normalization*, *Peak-to-Peak Normalization*, *Histogram Equalization*. Η εφαρμογή τους μπορεί να γίνει είτε καθολικά, με χρήση δηλαδή των πλαισίων του συνόλου των εκφωνήσεων, είτε ανά εκφώνηση, με χρήση δηλαδή των πλαισίων κάθε εκφώνησης ξεχωριστά. Επίσης, εφικτή είναι και η εφαρμογή κανονικοποίησης ανά ομιλητή, η οποία εξετάζεται στο Κεφάλαιο 6. Σε σύγκριση με την καθολική κανονικοποίηση, παρατηρήθηκε βελτίωση του ποσοστού επιτυχίας κατά την εφαρμογή των τεχνικών ανά εκφώνηση. Το αποτέλεσμα αυτό είναι λογικό, αφού τα δεδομένα μπορεί να προέρχονται από διαφορετικούς ομιλητές και ακουστικά περιβάλλοντα. Ωστόσο, δεν παρατηρήθηκε αξιοσημείωτη βελτίωση, συγκριτικά με την απόδοση του συγκεκριμένου συστήματος χωρίς κανονικοποίηση των δεδομένων. Στο κεφάλαιο αυτό, χρησιμοποιήθηκαν οι τέσσερις περιπτώσεις συστήματος με βάση το *Affective Saliency Model* [31].

Ιδιαίτερο ενδιαφέρον παρουσιάζουν μια σειρά από τεχνικές Προσαρμογής του Ομιλητή, οι οποίες έχουν αναπτυχθεί στον τομέα της Αυτόματης Αναγνώρισης Φωνής. Οι βασικότερες από αυτές είναι οι: *Cepstral Mean Normalization (CMN)*, *Vocal Tract Length Normalization (VTLN)*, *Maximum A-Posteriori Adaptation (MAP)*, *Maximum Likelihood Linear Regression (MLLR)* και *Speaker Adaptive Training (SAT)*. Οι τεχνικές αυτές επιδιώκουν την αύξηση της απόδοσης του μοντέλου κατά την αξιολόγηση των δεδομένων ενός ομιλητή, μέσω της προσαρμογής των χαρακτηριστικών του ή των παραμέτρων του μοντέλου. Ανάλυση αυτών των

τεχνικών παρατίθεται στο Κεφάλαιο 5, ενώ στο Κεφάλαιο 6 γίνεται εφαρμογή τους με σκοπό την Αναγνώριση Συναισθήματος από Φωνή. Με μοντέλο GMM, παρατηρήθηκε η χρησιμότητα των συγκεκριμένων τεχνικών στην αναγνώριση συναισθήματος, καθώς και η βελτίωση των αποτελεσμάτων κατά την εφαρμογή τους ανά ομιλητή, συγκριτικά με την εφαρμογή ανά εκφώνηση. Αξιοσημείωτη βελτίωση έδωσε η τεχνική προσαρμοζόμενης εκπαίδευσης SAT, η οποία χρησιμοποιεί τεχνική προσαρμογής και κατά την εκπαίδευση του συστήματος. Φαίνεται ότι δίνει καλύτερο διαχωρισμό των φωνητικών ιδιαιτεροτήτων των ομιλητών από τα καθαρά φωνητικά χαρακτηριστικά, τα οποία θα οδηγήσουν στην ορθή αναγνώριση συναισθήματος.

Σε συνέχεια των παραπάνω τεχνικών, επιχειρήθηκαν κάποιες παραλλαγές τους. Αρχικά, εξετάστηκε η περίπτωση προσαρμογής ανά συναισθηματική κλάση κάθε ομιλητή. Η συγκεκριμένη προσέγγιση, αν και δεν είναι ρεαλιστική με την άμεση μορφή της, έδειξε τη σημασία της διαφορετικής αντιμετώπισης κάθε συναισθηματικής κλάσης ξεχωριστά. Ενισχύει την πεποίθηση της διαφορετικής έκφρασης των ομιλητών. Επιπλέον, δίνει το κίνητρο ανάπτυξης συστημάτων, τα οποία θα υλοποιούν ξεχωριστή προσαρμογή των χαρακτηριστικών διαφορετικών κλάσεων. Διατηρώντας, όμως, την κλασσική προσαρμογή των χαρακτηριστικών ανά ομιλητή, δοκιμάστηκε η εκπαίδευση μοντέλου SVM με χρήση των μετασχηματισμένων χαρακτηριστικών. Επιβλήθηκε, δηλαδή, ο αφινικός μετασχηματισμός της τεχνικής fMLLR στα χαρακτηριστικά κάθε ομιλητή, όπως προκύπτει κανονικά, και στη συνέχεια, η ταξινόμηση έγινε με SVM. Η συγκεκριμένη προσέγγιση έδωσε σημαντική βελτίωση του ποσοστού επιτυχίας.

Με δεδομένη τη διαφορετική έκφραση των ομιλητών, μια άλλη προσέγγιση είναι η ανίχνευση της ουδέτερη ομιλίας τους και ο διαχωρισμός της από τη συναισθηματική. Σε αυτό το πνεύμα, στο [49] προτείνεται μια επαναληπτική διαδικασία κανονικοποίησης των χαρακτηριστικών. Σε κάθε επανάληψη, εφαρμόζεται κανονικοποίηση του συνόλου των δεδομένων, με βάση μόνο τα ουδέτερα χαρακτηριστικά που μόλις προβλέφθηκαν. Σκοπός αυτής της διαδικασίας είναι η ταύτιση των ουδέτερων χαρακτηριστικών μεταξύ των ομιλητών, με ταυτόχρονη διατήρηση της διακριτότητας των συναισθηματικών κλάσεων. Σημαντικό στοιχείο συνιστά η εφαρμογή της χωρίς επίβλεψη, κάνοντας εφικτή την υλοποίησή της σε πραγματικές εφαρμογές όπου ο ομιλητής είναι άγνωστος. Με βάση την παραπάνω ιδέα, αναπτύχθηκε παρόμοιο σύστημα όπου σε αντικατάσταση της τεχνικής κανονικοποίησης, εντάχθηκε η τεχνική προσαρμογής fMLLR, με σκοπό τον αφινικό μετασχηματισμό των χαρακτηριστικών. Οι πίνακες μετασχηματισμού εκτιμώνται με βάση μόνο τα ουδέτερα χαρακτηριστικά, τα οποία προβλέπονται σε κάθε επανάληψη. Το νέο σύστημα έδωσε σημαντική βελτίωση των αποτελεσμάτων, αξιοποιώντας επιπλέον την τεχνική SAT για την εκπαίδευση του μοντέλου GMM. Σημαντική συνεισφορά παρατηρήθηκε, επίσης, από την εφαρμογή κριτηρίου τερματισμού.

8.2 Μελλοντικές Προεκτάσεις

Σε συνέχεια της συγκεκριμένης εργασίας, προτείνονται παρακάτω πιθανές προεκτάσεις της. Στόχος είναι η έρευνα όλων των πιθανών στοιχείων, τα οποία θα οδηγήσουν σε βελτίωση των παραπάνω συστημάτων, καθώς και η ένταξη νέων σχετικών ιδεών. Αρχικά, ενδιαφέρουσα παράμετρος αποτελεί το σύνολο των εξαγόμενων χαρακτηριστικών. Στα παραπάνω συστήματα χρησιμοποιήθηκε το ίδιο σύνολο χαρακτηριστικών, με σκοπό την εύκολη σύγκριση. Με βάση τη σημαντική συμβολή της τεχνικής SAT με fMLLR, όπως παρατηρήθηκε σε όλες τις περιπτώσεις, προτείνεται η χρήση μεγαλύτερης ποικιλίας χαρακτηριστικών για την προσαρμογή

τους σε μοντέλο GMM. Απαιτείται η διερεύνηση πλήθους χαρακτηριστικών, όπως αυτά που αναφέρθηκαν στο Κεφάλαιο 2, και η επιλογή των καταλληλότερων, με σκοπό τη βελτίωση της τελικής απόδοσης του μοντέλου. Παραδείγματα χαρακτηριστικών που αξίζει να διερευνηθούν είναι αυτά που σχετίζονται με την ποιότητα της φωνής, όπως το jitter και το shimmer. Όπως αναφέρθηκε και στο Κεφάλαιο 2, η συναισθηματική κατάσταση του ομιλητή επηρεάζει σε κάποιο βαθμό την ποιότητα της φωνής του, κάτι που γίνεται αντιληπτό από τον άνθρωπο συνομιλητή του.

Αντίστοιχα, μελέτη των χαρακτηριστικών προτείνεται και στην περίπτωση ταξινόμησης με SVM, όπου εφαρμόζεται μετασχηματισμός fMLLR (βλ. Σχήμα 6.7). Υπενθυμίζεται ότι στο σύστημα αυτό εξάγονται ακουστικά χαρακτηριστικά ανά πλαίσιο, τα οποία προσαρμόζονται σε μοντέλο GMM με χρήση της τεχνικής fMLLR. Μετά την εφαρμογή των αφινικών μετασχηματισμών fMLLR ανά ομιλητή στα προηγούμενα χαρακτηριστικά, εξάγονται στατιστικά χαρακτηριστικά σε επίπεδο εκφώνησης, με σκοπό την SVM ταξινόμηση. Το συγκεκριμένο σύστημα, αποδείχτηκε ιδιαίτερα ενδιαφέρον, δίνοντας αξιοσημείωτη βελτίωση των αποτελεσμάτων. Προτείνεται, λοιπόν, η μελέτη τόσο των χαμηλού-επιπέδου χαρακτηριστικών, αφού η fMLLR είναι γενική ως τεχνική, όσο και των στατιστικών που υπολογίζονται για την εξαγωγή ανά εκφώνηση. Ως στατιστικά, εδώ, χρησιμοποιήθηκαν τα εξής 5: μέγιστη τιμή, ελάχιστη τιμή, μέση τιμή, διάμεσος και τυπική απόκλιση. Σε συνδυασμό με τα παραπάνω, είναι δυνατή η επιπλέον εξαγωγή πολλών άλλων στατιστικών (βλ. κεφ. 2), όπως: εκατοστημόρια, τεταρτημόρια, εύρος τιμών μεταξύ εκατοστημορίων, λοξότητα, κύρτωση, διάρκεια σήματος με πλάτος μεγαλύτερο/μικρότερο από κάποια τιμή ή με πλάτος που αυξάνεται/μειώνεται. Επίσης, μετά την προσθήκη τόσο των χαμηλού-επιπέδου χαρακτηριστικών όσο και των στατιστικών, μπορεί να εφαρμοστεί τεχνική επιλογής χαρακτηριστικών (*Feature Selection*) [55, 13, 49, 71], με σκοπό τη μείωση των διαστάσεων και τη βελτίωση της απόδοσης του συστήματος.

Παράλληλα με τη μελέτη της ποικιλίας των εξαγόμενων χαρακτηριστικών, για το παραπάνω σύστημα ταξινόμησης SVM, κρίνεται απαραίτητη και η μελέτη άλλων συναρτήσεων πυρήνα (kernel) εκτός από τη γραμμική, όπως η πολυωνυμική [30, 27] ή η RBF [55, 23, 52, 30, 27]. Αν και με αυτόν τον τρόπο, αυξάνεται η πολυπλοκότητα του συστήματος, η διαφορετική συνάρτηση απεικόνισης των νέων χαρακτηριστικών μπορεί να δώσει αποτελεσματικότερο διαχωρισμό των κλάσεων. Επιπλέον, είναι δυνατή η αντικατάσταση του ταξινομητή SVM με διαφορετικό μοντέλο, με σκοπό την ταξινόμηση των μετασχηματισμένων χαρακτηριστικών. Ιδιαίτερη ανάπτυξη παρατηρείται τα τελευταία χρόνια στα βαθιά νευρωνικά δίκτυα (Deep Neural Networks ή DNN), τα οποία έχουν αρχίσει να εφαρμόζονται και στην Αναγνώριση Συναισθήματος από Φωνή [28, 29, 74, 75]. Έτσι, θα ήταν ενδιαφέρουσα η χρήση τους στο σύστημα αυτό, με σκοπό την ταξινόμηση των εκφωνήσεων, μετά το μετασχηματισμό των χαρακτηριστικών με πίνακα fMLLR.

Όσον αφορά το σύστημα επαναληπτικής προσαρμογής των χαρακτηριστικών, με ανίχνευση των ουδέτερων εκφωνήσεων και διαχωρισμό τους από τις συναισθηματικές (βλ. Σχήμα 7.2), και σε αυτό προτείνεται εκτενέστερη διερεύνηση και επιλογή των ακουστικών χαρακτηριστικών. Επιπλέον, πιθανή βελτίωση μπορεί να δώσει κάποιο διαφορετικό κριτήριο τερματισμού των επαναλήψεων, οπότε και αυτό αποτελεί αντικείμενο μελέτης. Σχετικά με το μοντέλο για την ταξινόμηση των εκφωνήσεων, χρησιμοποιείται GMM, το οποίο έχει εκπαιδευτεί με SAT. Στα πειράματα που πραγματοποιήθηκαν, δοκιμάστηκε και η αντικατάσταση του μοντέλου με SVM με γραμμική συνάρτηση πυρήνα. Σε αυτήν την περίπτωση, χρησιμοποιήθηκαν τα μετασχηματισμένα χαρακτηριστικά με εφαρμογή πίνακα fMLLR, ακριβώς όπως στο προηγούμενο σύστημα του Κεφαλαίου 6. Η διαφορά εντοπίζεται στην υλοποίηση των επαναλήψεων, με δια-

χωρισμό των 2 κλάσεων χωρίς επίβλεψη. Η επιλογή αυτή, δηλαδή της SVM ταξινόμησης, δεν έδωσε κάποια βελτίωση μετά την πρώτη επανάληψη. Συμπερασματικά, πιθανή είναι η μελέτη άλλων συναρτήσεων πυρήνα, όπως παραπάνω, αλλά και άλλων μοντέλων, όπως κάποιο νευρωνικό δίκτυο.

Ενδιαφέρουσα, ακόμα, είναι η εξέλιξη του συγκεκριμένου συστήματος, με σκοπό την κατηγοριοποίηση περισσότερων κλάσεων, όπως οι 4 συναισθηματικές κλάσεις του Κεφαλαίου 6 (χαρά, λύπη, θυμός, ουδέτερο). Στην ένταξη περισσότερων συναισθηματικών κλάσεων, όπως φάνηκε τόσο από το Σχήμα 7.1 όσο και από σχετικό πείραμα, δε θα συμβάλλει η ίδια ιδέα, όπου ο πίνακας του αφινικού μετασχηματισμού εκτιμάται με βάση μόνο τις ουδέτερες εκφωνήσεις, που μόλις έχουν προβλεφθεί. Αυτό οφείλεται στη διαφορετική συναισθηματική έκφραση κάθε ομιλητή, όπως απεικονίζεται στο συγκεκριμένο σχήμα, ακόμα και μετά την ταύτιση της ουδέτερης ομιλίας των ομιλητών. Για την ένταξη περισσότερων συναισθηματικών κλάσεων, λοιπόν, είναι δυνατή η αξιοποίηση της ιδέας προσαρμογής ανά συναισθηματική κλάση του ομιλητή, όπως αναφέρθηκε στην παράγραφο 6.5. Σύμφωνα και με τα συμπεράσματα, η συγκεκριμένη ιδέα δείχνει τη σημασία της διαφορετικής αντιμετώπισης κάθε συναισθηματική κλάσης ξεχωριστά, ενισχύοντας την πεποίθηση της διαφορετικής έκφρασης μεταξύ των ομιλητών. Γενικά, η ιδέα προσαρμογής των χαρακτηριστικών κάθε κλάσης κάθε ομιλητή ξεχωριστά, είναι δυνατόν να αξιοποιηθεί κατά την ανάπτυξη πολυπλοκότερων συστημάτων, με ένταξη τεχνικών προσαρμογής.

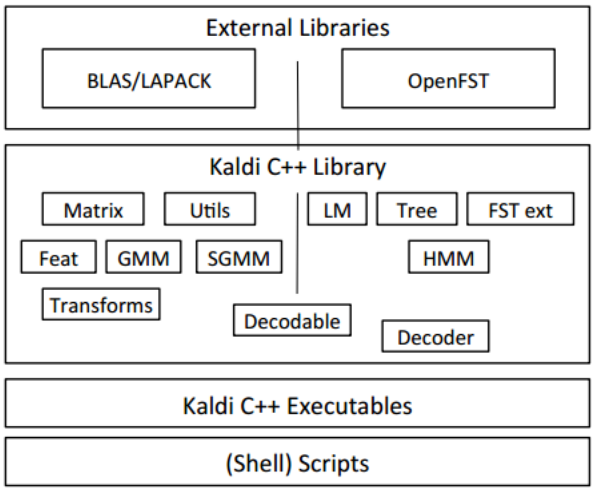
Όσον αφορά όλα τα παραπάνω συστήματα, είναι επιπλέον δυνατή η αξιολόγησή τους με χρήση διαφορετικής βάσης δεδομένων. Η βάση IEMOCAP είναι ιδιαίτερα πλούσια, καθώς περιλαμβάνει τόσο γραμμένα σενάρια όσο και αυτοσχεδιαστική ομιλία, προσεγγίζοντας έτσι πραγματικές καταστάσεις και συναισθήματα. Επίσης, ο χωρισμός leave-one-speaker-out καθιστά ακόμα πιο δύσκολη την αναγνώριση του συναισθήματος ενός καινούργιου ομιλητή. Ωστόσο, ενδιαφέρον θα έχει η αξιολόγηση των παραπάνω συστημάτων με διαφορετική βάση, σε σχέση με τα δεδομένα εκπαίδευσης, με σκοπό τη μελέτη της απόδοσης των τεχνικών προσαρμογής. Επιπλέον, ιδιαίτερο ενδιαφέρον έχει η συγκέντρωση πραγματικών δεδομένων, για παράδειγμα από τηλεοπτικές εκπομπές [49], και η χρήση τους για την αξιολόγηση κάθε συστήματος. Σε αυτή την περίπτωση, προβλέπεται μείωση της απόδοσης, καθώς η έκφραση συναισθημάτων στην πραγματικότητα ποικίλει και διαφοροποιείται ακόμα περισσότερο, ανάλογα με τους ομιλητές ή τις καταστάσεις. Ακόμα, θα απαιτείται επισημείωση των εκφωνήσεων, κάτι που είναι δύσκολο, καθώς συχνά παρατηρείται ασάφεια μεταξύ των συναισθημάτων και διαφορετική αντίληψη από άνθρωπο σε άνθρωπο (όπως αναφέρθηκε και στο Κεφάλαιο 1).

Τέλος, αξίζει να αναφερθεί ότι, σε όλες τις παραπάνω περιπτώσεις, θεωρήθηκε τέλειος διαχωρισμός ενός ομιλητή από τους υπόλοιπους για την αξιολόγηση κάθε συστήματος. Σε πραγματικές εφαρμογές, όμως, είναι πιθανή η ομιλία περισσότερων από έναν ομιλητή. Τότε, το σύστημα θα καλείται να αναγνωρίσει το συναίσθημα κάθε εκφώνησης, που μπορεί να προέρχεται από ποικιλία ομιλητών. Το ζητούμενο είναι αν το γεγονός αυτό επηρεάζει την απόδοσή του. Στο [49] (σύστημα επαναληπτικής κανονικοποίησης των χαρακτηριστικών), παρουσιάζεται ένα σχετικό πείραμα, το οποίο δείχνει μικρή πτώση του ποσοστού επιτυχίας, όταν στα δείγματα αξιολόγησης έχει γίνει λάθος αναγνώριση της ταυτότητας του ομιλητή μικρότερο του 50%. Με παρόμοιο τρόπο, εδώ, μπορεί να αξιολογηθεί η απόδοση όλων των συστημάτων. Επιπλέον, ενδιαφέρουσα κατεύθυνση είναι η ανάπτυξη μοντέλου αναγνώρισης του ομιλητή, το οποίο θα ενταχθεί σε κάθε σύστημα, πριν την όποια προσαρμογή των δεδομένων. Έτσι, θα γίνει εφικτός ο διαχωρισμός των διαφορετικών ομιλητών κατά την αξιολόγηση, με σκοπό την προσαρμογή των δεδομένων κάθε ομιλητή ξεχωριστά.

Παράρτημα Α

Kaldi

Το Kaldi [61] αποτελεί ένα εργαλείο αναγνώρισης φωνής, ανοιχτού κώδικα, γραμμένο σε γλώσσα C++ και με άδεια χρήσης από την Apache License v2.0. Αναπτύσσεται κυρίως από τους Daniel Povey κ.ά., με σκοπό τη χρήση του από ερευνητές αναγνώρισης φωνής. Ως ιδέα, ξεκίνησε το 2009 σε ένα εργαστήριο στο Johns Hopkins University. Το όνομά του προέρχεται από το μύθο, σύμφωνα με τον οποίο ο Kaldi ήταν ο Αιθίοπας γιδοβοσκός που ανακάλυψε το φυτό του καφέ.



Σχήμα Α.1: *Απλοποιημένο διάγραμμα των δομικών στοιχείων του Kaldi. (Η εικόνα προέρχεται από το [61].)*

Στο Σχήμα Α.1 απεικονίζεται ένα διάγραμμα των δομικών στοιχείων του Kaldi. Αρχικά, ως εξωτερικές βιβλιοθήκες χρησιμοποιεί τις εξής 2: OpenFst [76] και BLAS/LAPACK. Η OpenFst δίνει τη δυνατότητα δημιουργίας των Finite State Transducers (FSTs), τα οποία αντιστοιχούν σε αυτόματα, όπου κάθε μετάβαση έχει μια ετικέτα εισόδου, μία εξόδου, και πιθανώς και ένα βάρος. Χρησιμοποιούνται για την αναπαράσταση σειράς χαρακτήρων ή σχέσεων μεταξύ δύο τέτοιων σειρών. Αντίστοιχα, οι Basic Linear Algebra Subroutines (BLAS) και Linear Algebra PACKage (LAPACK) αποτελούν βιβλιοθήκες γραμμικής άλγεβρας. Με βάση τις 2

αυτές εξωτερικές βιβλιοθήκες, οι διαφορετικές ενότητες της βιβλιοθήκης του Kaldi μπορούν να χωριστούν σε 2 ξεχωριστά μέρη. Κάθε ένα από αυτά τα μέρη εξαρτάται από μία εξωτερική βιβλιοθήκη και ενώνεται με το άλλο με χρήση του Decodable Interface. Ουσιαστικά, κάθε λειτουργία παρέχεται μέσω εργαλείου στη γραμμή εντολών, γραμμένου σε C++. Κάθε τέτοιο εργαλείο μπορεί συμπεριληφθεί στη σύνταξη ενός script, με σκοπό την υλοποίηση ενός συστήματος αναγνώρισης φωνής.

Παρέχεται πλήθος εργαλείων, καλύπτοντας τις βασικότερες περιοχές της αναγνώρισης φωνής, όπως: η εξαγωγή χαρακτηριστικών, η εκπαίδευση ακουστικού και γλωσσικού μοντέλου, η προσαρμογή του ομιλητή, η κατασκευή δέντρων απόφασης και γράφων για αποκωδικοποίηση. Βασικό στοιχείο αποτελεί η διαθεσιμότητα πολλών επιλογών σε κάθε εργαλείο, με σκοπό την εύκολη τροποποίηση των προκαθορισμένων επιλογών από το χρήστη, αν αυτό είναι απαραίτητο. Επίσης, δίνεται έμφαση στην ανάπτυξη γενικών αλγορίθμων, έτσι ώστε να παρέχεται δυνατότητα επέκτασης.

Όσον αφορά την εξαγωγή χαρακτηριστικών, τα βασικότερα χαρακτηριστικά που χρησιμοποιούνται είναι τα MFCC και τα PLP. Κατά την εξαγωγή τους, δίνεται η δυνατότητα κανονικοποίησης ή μετασχηματισμού με τεχνικές όπως: CMVN, VTLN και LDA. Όσον αφορά τα ακουστικά μοντέλα, το Kaldi υποστηρίζει κυρίως την κατασκευή HMM (Hidden Markov Models), όπου οι συναρτήσεις πυκνότητας πιθανότητας κάθε φωνήματος εκτιμώνται με βάση μοντέλο GMM. Συνοπτικά, ένα HMM αντιστοιχεί σε μία αλυσίδα καταστάσεων, κάθε μία από τις οποίες αντιστοιχεί σε ένα φώνημα και παράγει μία παρατήρηση (διάνυσμα χαρακτηριστικών). Η ακολουθία των παρατηρήσεων θεωρείται γνωστή, ενώ η ακολουθία των καταστάσεων άγνωστη.

Εκτός από τα απλά GMM με διαγώνιο πίνακα συνδιακύμανσης, έχουν προστεθεί εργαλεία για ανάπτυξη πλήρους πίνακα συνδιακύμανσης. Επιπλέον, αξιοσημείωτη συμβολή του Kaldi αφορά τα Subspace Gaussian Mixture Models (SGMMs) [77], τα οποία όμως ξεπερνούν τα όρια της παρούσας μελέτης. Με βάση τα μοντέλα HMM/GMM, υποστηρίζονται και τεχνικές Προσαρμογής του Ομιλητή, με τροποποίηση είτε των παραμέτρων του μοντέλου (π.χ. MLLR) είτε των διανυσμάτων των χαρακτηριστικών (π.χ. fMLLR).

Κατά την εκπαίδευση ενός μοντέλου HMM/GMM, βασικό στοιχείο αποτελεί η δημιουργία δέντρων απόφασης. Με βάση ένα παράθυρο συμφραζομένων, ορίζονται ρίζες και ερωτήσεις που αφορούν τα φωνήματα μέσα σε αυτό το παράθυρο. Η δημιουργία τόσο του γλωσσικού μοντέλου, όσο και των γράφων για την αποκωδικοποίηση ενός HMM, βασίζεται στα FSTs. Ουσιαστικά, κατά την αποκωδικοποίηση κατασκευάζεται ένα πλέγμα (*lattice*), το οποίο αναπαριστά τις εναλλακτικές πιθανές ακολουθίες λέξεων για μία συγκεκριμένη εκφώνηση. Η πρόβλεψη της τελικής ακολουθίας γίνεται με μία αναδρομική διαδικασία κλαδέματος, η οποία βασίζεται στον αλγόριθμο Viterbi.

Επιπρόσθετες τεχνικές έχουν αναπτυχθεί, περιλαμβάνοντας MMI και MCE discriminative training, όπως επίσης και κώδικα για βαθιά νευρωνικά δίκτυα (DNNs). Στη συγκεκριμένη εργασία, δεν κρίνεται απαραίτητη η εκτενέστερη περιγραφή του συγκεκριμένου εργαλείου αναγνώρισης φωνής, καθώς οι βασικές λειτουργίες που αξιοποιούνται εδώ, αναλύονται στο Κεφάλαιο 5. Περισσότερες πληροφορίες παρέχονται στη ιστοσελίδα: http://kaldi-asr.org/doc/index.html.

Βιβλιογραφία

[1] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz και John G. Taylor. "Emotion recognition in human-computer interaction". *Signal Processing Magazine, IEEE* 18.1 (2001), σσ. 32–80.

[2] Rosalind W. Picard. "Affective Computing". *MIT press* 321 (1995), σσ. 1–16.

[3] Alexandros Potamianos. "Cognitive multimodal processing: from signal to behavior". *Proc. of Workshop on Roadmapping the Future of Multimodal Interaction Research* (2014).

[4] John Gideon, Emily Mower Provost και Melvin McInnis. "Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder". *ICASSP* (2016).

[5] Stefan Scherer, Gale M. Lucas, Jonathan Gratch, Albert Rizzo και Louis Philippe Morency. "Self-Reported Symptoms of Depression and PTSD Are Associated with Reduced Vowel Space in Screening Interviews". *IEEE Transactions on Affective Computing* 7.1 (2016), σσ. 59–73.

[6] Roddy Cowie. "Describing the emotional states expressed in speech". *ITRW on Speech and Emotion* (2000).

[7] Robert Plutchik. *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion*. New York: Academic, 1980.

[8] A. Mehrabian. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies*. Social Environmental and Developmental Studies. Oelgeschlager, Gunn & Hain, 1980.

[9] Hugo Lövheim. "A new three-dimensional model for emotions and monoamine neurotransmitters". *Med Hypotheses* 78 (2011), σσ. 341–348.

[10] Lawrence Rabiner και Ronald Schafer. *Theory and Applications of Digital Speech Processing*. 1st. Upper Saddle River, NJ, USA: Prentice Hall Press, 2011.

[11] Peter B. Denes και Elliot Pinson. *The Speech Chain*. Worth Publishers, 1993.

[12] Lawrence Rabiner και Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[13] Christos Nikolaos Anagnostopoulos, Theodoros Iliou και Ioannis Giannoukos. "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011". *Artificial Intelligence Review* 43.2 (2012), σσ. 155–177.

[14] Dimitrios Ververidis και Constantine Kotropoulos. "Emotional speech recognition: Resources, features, and methods". *Speech Communication* 48.9 (2006), σσ. 1162–1181.

[15] Eric Scheirer και Malcolm Slaney. "Construction and evaluation of a robust multifeature speech/music\ndiscriminator". *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing* 2 (1997), σσ. 2–5.

[16] Mireia Farrús, Javier Hernando και Pascual Ejarque. "Jitter and Shimmer Measurements for Speaker Recognition". *INTERSPEECH* (2007), σσ. 778–781.

[17] Carole T. Ferrand. "Harmonics-to-Noise Ratio". *Journal of Voice* 16.4 (2002), σσ. 480–487.

[18] Sungbok Lee, Serdar Yildirim, Abe Kazemzadeh και Shrikanth Narayanan. "An articulatory study of emotional speech production". *EUROSPEECH, Lisbon, Portugal* (2005).

[19] Sumi Shigeno. "Cultural similarities and differences in the recognition of audiovisual speech stimuli". *Proceedings of 5th International Conference on Spoken Language Processing* (1998), σσ. 149–152.

[20] James A. Russell. "Culture and the categorization of emotions". *Psychological bulletin* 110.3 (1991), σσ. 426–450.

[21] Batja Mesquita και Nico Frijda. "Cultural Variations in Emotions - A Review". *Phychological Bulletin* 112 (1992), σσ. 179–204.

[22] Kazuhito Koike, Hirotaka Suzuki και Hiroaki Saito. "Prosodic Parameters in Emotional Speech". *Proceedings of ICSLP 1998* (1998).

[23] Björn Schuller, Gerhard Rigoll και Manfred Lang. "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - belief network architecture". *Acoustics, Speech, and Signal Processing* 1 (2004), σσ. 577–580.

[24] Pierre Dumouchel, Najim Dehak, Yazid Attabi, Reda Dehak και Narjès Boufaden. "Cepstral and long-term features for emotion recognition". *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2 (2009), σσ. 344–347.

[25] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee και Shrikanth Narayanan. "Emotion recognition using a hierarchical binary decision tree approach". *Speech Communication* 53.9-10 (2011), σσ. 1162–1171.

[26] Yixiong Pan, Peipei Shen και Liping Shen. "Speech emotion recognition using support vector machine". *International Journal of Smart Home* 6.2 (2012), σσ. 101–108.

[27] Thapanee Seehapoch και Sartra Wongthanavasu. "Speech Emotion Recognition Using Support Vector Machines". *2013 5th International Conference on Knowledge and Smart Technology (KST)* (2013), σσ. 86–91.

[28] Andre Stuhlsatz, Christine Meyer, Florian Eyben, Thomas Zielke, Günter Meier και Björn Schuller. "Deep Neural Networks for Acoustic Emotion Recogition: Raising the Benchmarks". *IEEE ICASSP 2011* (2011), σσ. 5688–5691.

[29] Kun Han, Dong Yu και Ivan Tashev. "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine". *INTERSPEECH* (2014), σσ. 223–227.

[30] Hao Hu, Ming-Xing Xu και Wei Wu. "GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition". *ICASSP Proceedings* April (2007), σσ. 413–416.

[31] Arodami Chorianopoulou, Polychronis Koutsakis και Alexandros Potamianos. "Speech Emotion Recognition Using Affective Saliency". *Interspeech 2016* (2016), σσ. 500–504.

[32] Richard O. Duda, Peter E. Hart και David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[33] Sergios Theodoridis και Konstantinos Koutroumbas. *Pattern Recognition, Fourth Edition*. 4th. Academic Press, 2012.

[34] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[35] URL: `http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html`.

[36] URL: `https://www.linkedin.com/pulse/support-vector-machine-srinivas-kulkarni`.

[37] Yariv Ephraim, Amir Dembo και Lawrence R. Rabiner. "A Minimum Discrimination Information Approach for Hidden Markov Modeling". *IEEE Transactions on Information Theory,* 35 (1989), σσ. 1001–1013.

[38] Biing Hwang Juang και Shigeru Katagiri. "Discriminative Learning for Minimum Error Classification". *IEEE Transactions on Signal Processing* 40 (1992), σσ. 3043–3054.

[39] Zheng-Hua Tan και Børge Lindberg. "Low-Complexity Variable Frame Rate Analysis for Speech Recognition and Voice Activity Detection". *IEEE Journal of Selected Topics in Signal Processing* 4 (2010), σσ. 798–807.

[40] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann και Ian H. Witten. "The WEKA Data Mining Software: An Update". *SIGKDD Explorations* 11 (2009).

[41] MATLAB. *version R2015a*. Natick, Massachusetts: The MathWorks Inc., 2015.

[42] Tong Zhang, Mark Hasegawa-Johnson και Stephen E. Levinson. "Mental state detection of dialogue system users via spoken language". *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* (2003).

[43] Xuang Hung Le, Georges Quénot και Eric Castelli. "Recognizing emotions for the audio-visual document indexing". *Ninth Int'l Symp. Computers and Communications (ISCC '04)* 2 (2004), σσ. 580–584.

[44] Martin Wöllmer, Florian Eyben, Stephan Reiter, Björn Schuller, Cate Cox, Ellen Douglas-Cowie και Roddy Cowie. "Abandoning emotion classes - Towards continuous emotion recognition with modelling of long-range dependencies". *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (2008), σσ. 597–600.

[45] C. Clavel, I. Vasilescu, L. Devillers, G. Richard και T. Ehrette. "Fear-type emotion recognition for future audio-based surveillance systems". *Speech Communication* 50 (2008), σσ. 487–503.

[46] Dmitri Bitouk, Ragini Verma και Ani Nenkovab. "Class-Level Spectral Features for Emotion Recognition". *Speech Communication* 52 (2010), σσ. 613–625.

[47] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee και Shrikanth Narayanan. "Emotion recognition using a hierarchical binary decision tree approach". *Speech Communication* 53 (2011), σσ. 1162–1171.

[48] Angeliki Metallinou, Athanasios Katsamanis και Shrikanth Narayanan. "A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs". *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (2012), σσ. 2401–2404.

[49] Carlos Busso, Soroosh Mariooryad, Angeliki Metallinou και Shrikanth Narayanan. "Iterative feature normalization scheme for automatic emotion detection from speech". *IEEE Transactions on Affective Computing* 4.4 (2013), σσ. 386–397.

[50] Björn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth και Gerhard Rigoll. "Cross-corpus acoustic emotion recognition: Variances and strategies". *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015* 1 (2015), σσ. 470–476.

[51] URL: `https://en.wikipedia.org/wiki/Normal_distribution`.

[52] Tsang-Long Pao, Charles S. Chien, Yu-Te Chen, Jun-Heng Yeh, Yun-Maw Cheng και Wen-Yuan Liao. "Combination of Multiple Classifiers for Improving Emotion Recognition in Mandarin Speech". *Third Int'l Conf. Intelligent Information Hiding and Multimedia Signal Processing* 1 (2007), σσ. 35–38.

[53] Bogdan Vlasenko, Björn Schuller, Kinfe Tadesse Mengistu, Gerhard Rigoll και Andreas Wendemuth. "Balancing spoken content adaptation and unit length in the recognition of emotion and interest". *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (2008), σσ. 805–808.

[54] Ali Hassan, Robert Damper και Mahesan Niranjan. "On acoustic emotion recognition: Compensating for covariate shift". *IEEE Transactions on Audio, Speech and Language Processing* 21.7 (2013), σσ. 1458–1468.

[55] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao και Te-Won Lee. "Emotion Recognition by Speech Signals". *Eighth European Conf. Speech Comm. and Technology (EUROSPEECH 2003)* (2003), σσ. 125–128.

[56] Kelsey Ramírez-Gutiérrez, Daniel Cruz-Pérez και Héctor Pérez-Meana. "Face Recognition and Verification using Histogram Equalization.pdf". *ACS'10 Proceedings of the 10th WSEAS international conference on Applied computer science* 1 (2010), σσ. 85–89.

[57] Angel De La Torre, Antonio M. Peinado, Jose C. Segura, Jose L. Perez-Cordoba, Ma Carmen Benitez και Antonio J. Rubio. "Histogram equalization of speech representation for robust speech recognition". *IEEE Transactions on Speech and Audio Processing* 13 (2005), σσ. 355–366.

[58] Martin Wollmer, Erik Marchi, Stefano Squartini και Bjorn Schuller. "Multi-stream LSTM-HMM decoding and histogram equalization for noise robust keyword spotting". *Cognitive Neurodynamics* 5 (2011), σσ. 253–264.

[59] Alexander Schmitt, Stefan Ultes και Wolfgang Minker. "A parameterized and annotated spoken dialog corpus of the CMU Let's Go bus information system". *Proc. LREC 2012* (2012), σσ. 3369–3373.

[60] Florian Eyben, Felix Weninger, Florian Gross και Björn Schuller. "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor". *Proc. of ACM Multimedia 2013* (2013), σσ. 835–838.

[61] Daniel Povey κ.ά. "The Kaldi Speech Recognition Toolkit". *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding* (2011).

[62] Michael Pitz και Hermann Ney. "Vocal Tract Normalization Equals Linear Transformation in Cepstral Space". *IEEE Transactions on Speech and Audio Processing* 13.5 (2005), σσ. 930–944.

[63] Daniel Povey, Geoffrey Zweig και Alex Acero. "Speaker Adaptation with an Exponential Transform". *IEEE Workshop Automatic Speech Recognition & Understanding (ASRU)* (2011).

[64] URL: `http://kaldi-asr.org/doc/transform.html`.

[65] Juri Ganitkevitch. "Speaker Adaptation using Maximum Likelihood Linear Regression". *Rheinisch-Westfälische Technische Hochschule Aachen Lehrstuhl für Informatik VI - Seminar Automatic Speech Recognition* (2005).

[66] Daniel Povey και George Saon. "Feature and model space speaker adaptation with full covariance Gaussians". *ICSLP* (2006).

[67] Jae-Bok Kim, Jeong-Sik Park και Yung-Hwan Oh. "On-line speaker adaptation based emotion recognition using incremental emotional information". *ICASSP 2011* (2011), σσ. 4948–4951.

[68] Jianbo Jiang, Zhiyong Wu, Mingxing Xu, Jia Jia και Lianhong Cai. "Comparison of adaptation methods for GMM-SVM based speech emotion recognition". *IEEE Workshop on Spoken Language Technology, SLT 2012 - Proceedings* (2012), σσ. 269–273.

[69] Jun Deng, Zixing Zhang, Florian Eyben και Björn Schuller. "Autoencoder-based unsupervised domain adaptation for speech emotion recognition". *IEEE Signal Processing Letters* 21.9 (2014), σσ. 1068–1072.

[70] Qirong Mao, Wentao Xue, Qiru Rao, Feifei Zhang και Yongzhao Zhan. "Domain adaptation for speech emotion recognition by sharing priors between related source and target classes". *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2016-May (2016), σσ. 2608–2612.

[71] Mohammed Abdelwahab και Carlos Busso. "Supervised Domain Adaptation for Emotion Recognition from Speech". *ICASSP* (2015), σσ. 5058–5062.

[72] Carlos Busso, Murtaza Bulut, Chi-chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee και Shrikanth S. Narayanan. "IEMOCAP: Interactive emotional dyadic motion capture database". *Journal of Language Resources and Evaluation* 52.4 (2008), σσ. 335–359.

[73] Jae-bok Kim και Jeong-sik Park. "Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition". *Engineering Applications of Artificial Intelligence* 52 (2016), σσ. 126–134.

[74] Rui Xia και Yang Liu. "Leveraging Valence and Activation Information via Multitask Learning for Categorical Emotion Recognition". *IEEE ICASSP 2015* (2015), σσ. 5301–5305.

[75] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency και Stefan Scherer. "Representation Learning for Speech Emotion Recognition". *INTERSPEECH 2016* (2016), σσ. 3603–3607.

[76] Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut και Mehryar Mohri. "OpenFst: A General and Efficient Weighted Finite-State Transducer Library". *Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA 2007)*. Lecture Notes in Computer Science 4783 (2007), σσ. 11–23.

[77] Daniel Povey κ.ά. "The Subspace Gaussian Mixture model-A Structured Model for Speech Recognition". *Computer Speech & Language* 25.2 (2011), σσ. 404–439.