Πανεπιστημιο Θεσσαλιας

Πολυτεχνικη Σχολη

Τμημα Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων

Μεθοδοι βαθιας μαθησης για αναγνωριση συναισθηματων με βαση ηχο

Deep learning methods for audio-based emotion recognition

Διπλωματικη Εργασια

Κυρτσογλου Αλκιβιαδης

Επιβλεποντες Καθηγητες:

Αν. Καθ. Ποταμιανος Γερασιμος

Επικ. Καθ. Αργυριου Αντωνιος

Αν. Καθ. Βασιλακοπουλος Μιχαηλ

Βολος, Φεβρουαριος 2020

Περίληψη

Όσο η τεχνολογία εξελίσσεται, η επικοινωνία ανθρώπου-μηχανής έχει φτάσει στο σημείο όπου οι ηλεκτρονικοί υπολογιστές είναι ικανοί πλέον να αναγνωρίζουν τα ανθρώπινα συναισθήματα. Η διαδικασία αυτή, γνωστή ως αναγνώριση συναισθήματος, απασχολεί πολλές έρευνες σε διαφορετικά επιστημονικά πεδία τα οποία σχετίζονται με τις επιστήμες της πληροφορικής και της επικοινωνίας ανθρώπου-μηχανής. Με βάση τον τρόπο που επικοινωνούνται τα συναισθήματα, η αναγνώριση συναισθημάτων εξετάζεται από δύο σκοπιές: την οπτική επικοινωνία (π.χ εκφράσεις προσώπου) και προφορική επικοινωνία. Η παρούσα πτυχιακή επικεντρώνεται στην αναγνώριση συναισθήματος που προέρχεται από τον ήχο. Οι τρεις βασικοί πυλώνες της εργασίας είναι: η εύρεση δεδομένων και η επεξεργασία τους, η εξαγωγή κατάλληλων ηχητικών χαρακτηριστικών και η εφαρμογή ταξινομητών.

Η βάση δεδομένων που χρησιμοποιείται είναι, η "The Interactive emotional dyadic motion capture database (IEMOCAP)" και περιέχει ομιλίες μεταξύ ηθοποιών οι οποίοι προσομοιώνουν συζητήσεις με συναισθηματική φόρτιση. Από αυτές εξάγονται χαρακτηριστικά τα οποία περικλείουν πληροφορία συναισθήματος. Κατά τη διαδικασία της ταξινόμησης, τα χαρακτηριστικά χρησιμοποιούνται ώστε να εντοπίσουν το συναίσθημα κάθε έκφρασης. Εστιάζουμε σε δύο είδη ταξινομητών οι οποίοι έχουν ως βάση τους τα Κρυφά Μοντέλα Markov. Ο πρώτος κάνει χρήση Γκαουσιανών Μοντέλων Μίξης για τη μοντελοποίηση της κατανομής των πιθανοτήτων για κάθε κλάση συναισθήματος, ενώ ο δεύτερος χρησιμοποιεί βαθιά Νευρωνικά Δίκτυα. Για τη δημιουργία τους χρησιμοποιείται το εργαλείο Kaldi.

Abstract

As technology advances, human-machine interaction is improving noticeably to the extent that computers can identify human emotion. This process, also known as emotion recognition, is the major focus in recent research in different disciplines related to information sciences and Human-Computer Interaction. Based on the means emotions are communicated, emotion recognition is examined based on two aspects: visual communication (i.e., facial expressions) and verbal communication. This thesis focuses on the recognition of human emotions, emanating from speech audio. The research is concentrated on three main sections: data acquisition and processing, extraction of suitable audio features and the implementation of classifiers.

The Interactive emotional dyadic motion capture database (IEMOCAP) is the dataset we use in this research and contains speech utterances of actors who simulate different emotion conditions in a conversation. From these utterances we extract different combinations of features which best describe the emotion information. The classification procedure uses these features to recognize the emotion in each utterance. We primarily emphasize on two classifiers which both train Hidden Markov Models. The first model makes use of Gaussian Mixture Models (GMMs) to model the class conditional probabilities, while the other employs Deep Neural Networks (DNNs). The tool used in this procedure is the Kaldi toolkit.

Ευχαριστίες

Σε αυτό το σημείο θα ήθελα να εκφράσω τις ευχαριστίες σε όσους συνέβαλαν ώστε να ολοκληρωθεί η εργασία και το ταξίδι αυτό. Κατ' αρχάς, θα ήθελα να ευχαριστήσω τον βασικό επιβλέποντα καθηγητή μου, τον Αν. Καθ. Ποταμιάνο Γεράσιμο που μου έδωσε την ευκαιρία να εργαστώ πάνω σε έναν τόσο ενδιαφέρον τομέα και την υπομονή και καθοδήγηση που μου έδειξε όλο αυτό τον καιρό. Επίσης, θα ήθελα να ευχαριστήσω τα υπόλοιπα μέλη της επιτροπής επίβλεψης, τον Επικ. Καθ. Αντώνιο Αργυρίου και τον Αν. Καθ. Βασιλακόπουλο Μιχαήλ, για τον χρόνο που διέθεσαν για την ανάγνωση και την καθοδήγησή τους. Επίσης, ευχαριστώ από την καρδιά μου την οικογένειά μου για την στήριξή τους όλα αυτά τα χρόνια των σπουδών μου και όχι μόνο. Τέλος, θέλω να πω ένα ευχαριστώ στους συμφοιτητές μου, στο Κοχύλι, για τα ωραιότερα φοιτητικά χρόνια που μου χάρισαν και φυσικά στο αστέρι μου.

Στην οικογένειά μου...

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κεφάλαιο 1

Εισαγωγή

Το πρώτο βήμα για την διερεύνηση της έννοιας της αναγνώρισης συναισθήματος είναι να απαντήσουμε στην ερώτηση "τι είναι συναίσθημα;". Αυτό το ερώτημα απασχόλησε πολλά λαμπρά μυαλά της επιστήμης, διαφορετικών επιστημονικών υπόβαθρων, και αποτελεί θέμα μελέτης και έρευνας από τα αρχαία χρόνια. Στα [1] και [2] περιγράφονται με λεπτομέρεια ορισμένες θεωρίες συναισθημάτων. Για παράδειγμα, ο Αριστοτέλης πίστευε ότι το να έχεις κάποιο συναίσθημα σημαίνει να βιώνεις πόνο, ευχαρίστηση ή και τα δύο. Μεγάλη ήταν η συνεισφορά του Δαρβίνου στην μελέτη συναισθηματικών εκφράσεων, ο οποίος με βάση τη θεωρία της εξέλιξης υποστήριζε ότι τα συναισθήματά μας προέρχονται από πρόδρομα είδη. Επιπλέον, μια ακόμη θεμελιώδης θεωρία του Δαρβίνου την οποία ενστερνιζόταν και ο Descartes αλλά και μεταγενέστεροι επιστήμονες, όπως ο Tomkins και ο Ekman, είναι αυτή της ύπαρξης ενός διακριτού αριθμού βασικών συναισθημάτων.

Θεωρίες σαν τις προαναφερθείσες αποτελούν βασικό δομικό συστατικό σε πολλές έρευνες που γίνονται σήμερα στο πεδίο της αναγνώρισης συναισθήματος. Αυτό αποδεικνύει πόσο σημαντική είναι η γνώση αυτή για τη δημιουργία εκλεπτυσμένων συστημάτων αναγνώρισης. Ένα παράδειγμα φαίνεται στο [3], όπου εξετάζεται η αναγνώριση της ομιλίας με βάση την οπτική των βασικών συναισθημάτων του Ekman ενώ το [4], αναφέρεται

σε μια εναλλακτική προσέγγιση των συναισθημάτων σε ένα δισδιάστατο χώρο (βλ. Κεφάλαιο 2).

Η δημιουργία τέτοιων συστημάτων αναγνώρισης συναισθημάτων μπορεί να βοηθήσει στην επίλυση προβλημάτων της καθημερινότητας. Σε ιατρικές περιπτώσεις που η καταγραφή της συναισθηματικής κατάστασης του ασθενούς είναι απαραίτητη, μπορούν να συμβάλλουν στην μεγιστοποίηση της καταγραφής έγκυρων μετρήσεων. Επίσης, μεγάλη εφαρμογή έχουν και σε υπηρεσίες εξυπηρέτησης πελατών και τηλεφωνικών κέντρων [5], όπου η ανίχνευση της συναισθηματικής κατάστασης των πελατών μπορεί να συντελέσει στην καλύτερη δυνατή εμπειρία για αυτούς. Άλλες εφαρμογές μπορεί να αφορούν σε αυτοκίνητα που μπορούν να κατανοήσουν τα επίπεδα άγχους του οδηγού, ανίχνευση ψεύδους, αλλά ακόμη και στην διασκέδαση με ηλεκτρονικά παιχνίδια τα οποία μπορούν να κατανοήσουν ερεθίσματα έκφρασης από τον χρήστη τους.

1.1 Η προσέγγισή μας

Η εργασία αυτή, είναι μια προσπάθεια υλοποίησης ενός συστήματος αναγνώρισης συναισθημάτων από ηχητικά αποσπάσματα ομιλίας. Στόχος είναι η χρησιμοποίηση βαθιών νευρωνικών δικτύων (DNN) και ο έλεγχος της αποτελεσματικότητάς τους μέσα σε ένα τέτοιο περιβάλλον. Επιπλέον, εξετάζονται όλα τα απαραίτητα βήματα προεργασίας που πρέπει να γίνουν για την καλύτερη δυνατή απόδοση του συστήματος, στα οποία περιλαμβάνονται το είδος των δεδομένων που χρησιμοποιείται και η κατάλληλη επεξεργασία τους.

Η συνεισφορά της εργασίας αυτής μπορεί να συνοψιστεί στα εξής σημεία:

- Σύγκριση απόδοσης διαφορετικών δομών μοντέλων αναγνώρισης συναισθήματος από αποσπάσματα ομιλιών.
- Αναζήτηση για το ποια χαρακτηριστικά περιγράφουν τις συναισθηματικές εναλλαγές στον ήχο.

- Παράδειγμα χρήσης του εργαλείου Kaldi για αναγνώριση συναισθήματος.

1.2 Σχετική Έρευνα

Οι έρευνες που έχουν ασχοληθεί με το αντικείμενο που πραγματεύεται η εργασία αυτή είναι πάρα πολλές. Ένας λόγος είναι ότι τα νευρωνικά δίκτυα επιδέχονται πειραματισμούς λόγω των πολλών παραμέτρων τους και των διαφορετικών δομών που συναντώνται σε αυτά. Στην [6], γίνεται χρήση DNN τα οποία εξάγουν μια κατανομή πιθανοτήτων για τις καταστάσεις των συναισθημάτων. Στη συνέχεια, μαζί με τα χαρακτηριστικά που προσδιορίζουν κάθε έκφραση, μπαίνουν ως είσοδος σε μια δομή που την αναφέρουν ως "extreme learning machine (ELM)" που είναι στην ουσία ένα νευρωνικό δίκτυο με ένα κρυφό στρώμα. Με αυτό τον τρόπο κατάφεραν να βελτιώσουν τον εντοπισμό χρήσιμης πληροφορίας κατά 20% σε σχέση με σχετικές έρευνες.

Μία άλλη έρευνα [7], χρησιμοποιώντας το Kaldi για την δημιουργία των μοντέλων νευρωνικών δικτύων, αναζητά τρόπους βελτίωσης οι οποίοι βασίζονται στην εξαγωγή ενός είδους χαρακτηριστικών, τα epoch, τα οποία προσδιορίζουν τις περιοχές, μέσα σε ένα ηχητικό απόσπασμα, στις οποίες οι φωνητικές χορδές διεγείρονται σημαντικά. Το συμπέρασμα είναι πως η μεμονωμένη χρήση τους αποδίδει χειρότερα σε σχέση με τα πειράματα όπου συνδύασαν τα χαρακτηριστικά αυτά με άλλα όπως τα MFCCs για τα οποία θα μιλήσουμε στο Κεφάλαιο 3).

Τέλος, στην [8] εξετάζονται ταξινομητές CNN που έχουν ως είσοδο χαρακτηριστικά MFCC. Επιπλέον εξετάζεται ο συνδυασμός τους με γραπτές μεταγραφές, παρουσιάζοντας έτσι την πτυχή του συνδυασμού πληροφορίες από διαφορετικές πηγές.

1.3 Δομή της πτυχιακής

Η εργασία αυτή χωρίζεται σε 6 Κεφάλαια. Αναλυτικά:

- **Κεφάλαιο 2**, περιλαμβάνει πληροφορίες για τις ιδιότητες των δεδομένων που χρησιμοποιούνται σε αυτόν τον κλάδο της αναγνώρισης συναισθημάτων. Γίνεται περιγραφή των βάσεων δεδομένων που υπάρχουν, καθώς επίσης και για τη βάση που χρησιμοποιήθηκε κατά κύριο λόγο σε αυτή την εργασία.

- **Κεφάλαιο 3**, όπου γίνεται μια εισαγωγή της έννοιας των ακουστικών χαρακτηριστικών. Στη συνέχεια, γίνεται περιγραφή γνωστών χαρακτηριστικών που συναντώνται στη βιβλιογραφία και που αποδίδουν καλύτερα τη συναισθηματική κατάσταση του ομιλητή, όπως και τα εργαλεία που χρησιμοποιήθηκαν για την εξαγωγή και επεξεργασία τους.

- **Κεφάλαιο 4**, περιγράφει τη βασική θεωρία πίσω από τους ταξινομητές που χρησιμοποιούνται στον τομέα της αναγνώρισης συναισθήματος ομιλίας. Επιπλέον, γίνεται επεξήγηση του βασικού εργαλείου Kaldi, και καταλήγει στην περιγραφή των μοντέλων αναγνώρισης που χρησιμοποιήθηκαν στην εργασία.

- **Κεφάλαιο 5**, είναι το κεφάλαιο όπου παρατίθενται τα πειράματα που έγιναν με τους δύο ταξινομητές (GNN-HMM και DNN-HMM) και τα αποτελέσματα που προέκυψαν αντίστοιχα.

- **Κεφάλαιο 6**, το τελευταίο κεφάλαιο στο οποίο γίνεται η ανακεφαλαίωση της πτυχιακής και προτείνονται μελλοντικές ενέργειες για τη βελτίωση των αποτελεσμάτων.

Κεφάλαιο 2

Δεδομένα

Το πρώτο βήμα για την δημιουργία μηχανών με τη δυνατότητα να ξεχωρίζουν και να κατανοούν συναισθήματα, είναι η συλλογή κατάλληλων δεδομένων με τα οποία θα μπορέσουν να εκπαιδευτούν, αλλά και να αξιολογηθούν. Για αυτό το λόγο, ο ρόλος μιας βάσης δεδομένων στην υλοποίηση συστημάτων ανίχνευσης συναισθημάτων είναι θεμελιώδης.

Σύμφωνα με τη βιβλιογραφία ([9], [10]), οι βάσεις δεδομένων μπορούν να διαχωριστούν σε κατηγορίες ανάλογα με τις ιδιότητες που διαθέτουν τα περιεχόμενά τους.

- Με κριτήριο το είδος της ομιλίας προκύπτουν τρεις κατηγορίες: Δεδομένα βασισμένα σε φυσική ομιλία, προσομοιωμένη ομιλία και εκμαιευμένη ομιλία. Στη φυσική ομιλία, οι διάλογοι είναι αυθόρμητοι και τα συναισθήματα αποτυπώνονται με αληθοφάνεια. Η ομιλία που προκύπτει από σενάρια προσομοίωσης διαλόγων γίνεται συνήθως με ηθοποιούς που υποδύονται συναισθηματικά φορτισμένους ρόλους. Τέλος, όπως υποδηλώνει και το όνομα, στην εκμαιευμένη ομιλία τα συναισθήματα αποτυπώνονται μέσα από μια προσομοίωση συναισθηματικής κατάστασης για την οποία δεν έχει γνώση ο ομιλητής. Αυτό στοχεύει στην πιο αυθόρμητη καταγραφή των αντιδράσεών του.

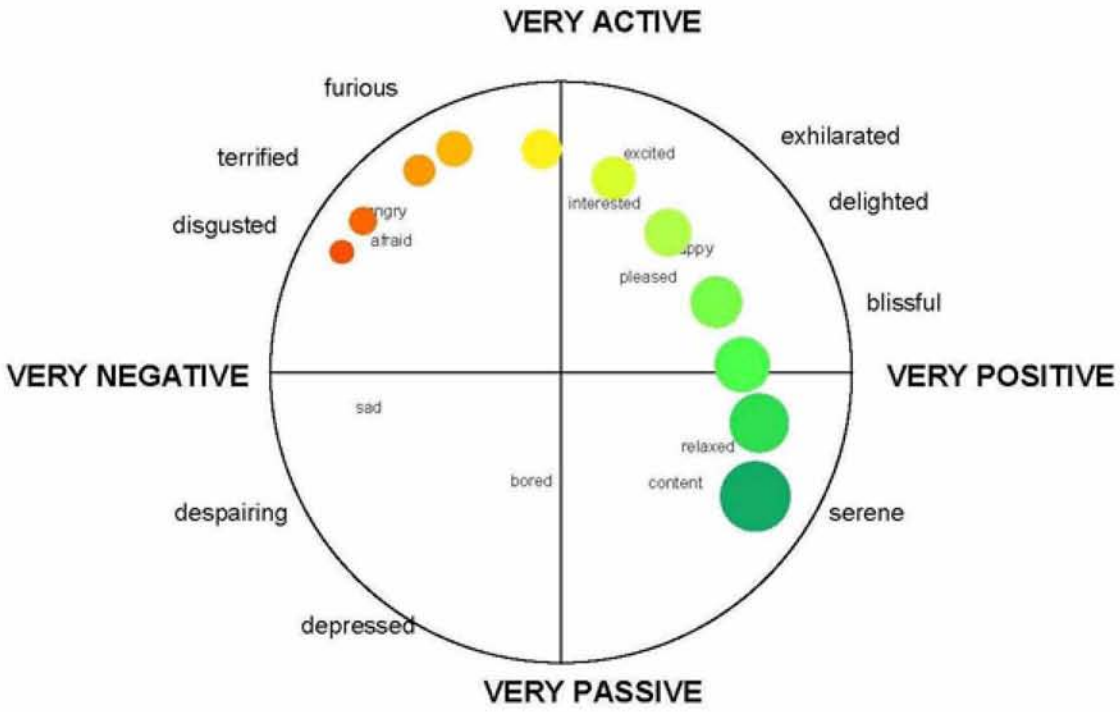
Οι πιο διαδεδομένες κατηγορίες είναι οι δύο τελευταίες, καθώς είναι πιο πρακτικό και εύκολο να δημιουργηθούν οι συνθήκες που απαιτούνται για την απόκτηση τέτοιων δεδομένων. Ωστόσο, υπάρχουν κάποια σημαντικά μειονεκτήματα [11] που τις συνοδεύουν και είναι σημαντικό να αναφερθούν.

Πρώτον, τα επιτηδευμένα συναισθήματα δεν αποδίδονται όπως τα αυθόρμητα. Οι ηθοποιοί τείνουν ορισμένες φορές να υπερβάλλουν αποδίδοντας έτσι μια πιο αφύσικη αποτύπωσή τους.

Επιπλέον, οι συνθήκες μέσα στις οποίες γίνονται οι ηχογραφήσεις δεν αφήνουν πολλά περιθώρια για φυσικότητα. Γνωρίζοντας κάποιος ότι ηχογραφείται δεν μπορεί να αποδώσει σε καμία περίπτωση τις φυσικές του αντιδράσεις. Το ίδιο ισχύει και όταν καλείται να μιλήσει μόνος του χωρίς κάποιο άλλο άτομο στην συζήτηση.

Ένα άλλο μειονέκτημα είναι η καταγραφή και η αξιολόγηση των συναισθημάτων [12]. Με άλλα λόγια, πόσο αντικειμενικά μπορεί κάποιος να συμπεράνει το συναίσθημα που περιέχεται σε ένα δείγμα ομιλίας. Η διαδικασία αυτή είναι πολύ υποκειμενική και για αυτό το λόγο οι περισσότερες έρευνες πάνω στη δημιουργία τέτοιων βάσεων χρησιμοποιούν τουλάχιστον 2 αξιολογητές.

- Με κριτήριο την κατηγοριοποίηση των συναισθημάτων. Δύο είδη συναντάμε σε αυτή την κατηγορία, τις βάσεις που χρησιμοποιούν διακριτές ετικέτες για τις κλάσεις των συναισθημάτων και αυτές που απεικονίζουν τα συναισθήματα σε ένα δισδιάστατο χώρο. Στην πρώτη περίπτωση κάθε απόσπασμα ομιλίας ανήκει σε μία από τις υπάρχουσες κλάσεις μετά από τη διαδικασία αξιολόγησης. Στην δεύτερη περίπτωση, ένα απόσπασμα ομιλίας περιγράφεται από δύο τιμές, σθένους και διέγερσης. Η περιγραφή αυτή τοποθετεί κάθε δείγμα σε ένα δισδίαστατο χώρο απεικόνισης συναισθημάτων (Εικόνα 2.1). Με αυτό τον τρόπο οι ερευνητές προσδοκούν μεγαλύτερη ακρίβεια στην καταγραφή του πραγματικού συναισθήματος χωρίς να περιορίζονται από αυστηρά ορισμένες κλάσεις.



Σχήμα 2.1: Δυσδιάστατη απεικόνιση συναισθημάτων. Εικόνα από [13].

2.1 Βάσεις δεδομένων που χρησιμοποιήθηκαν

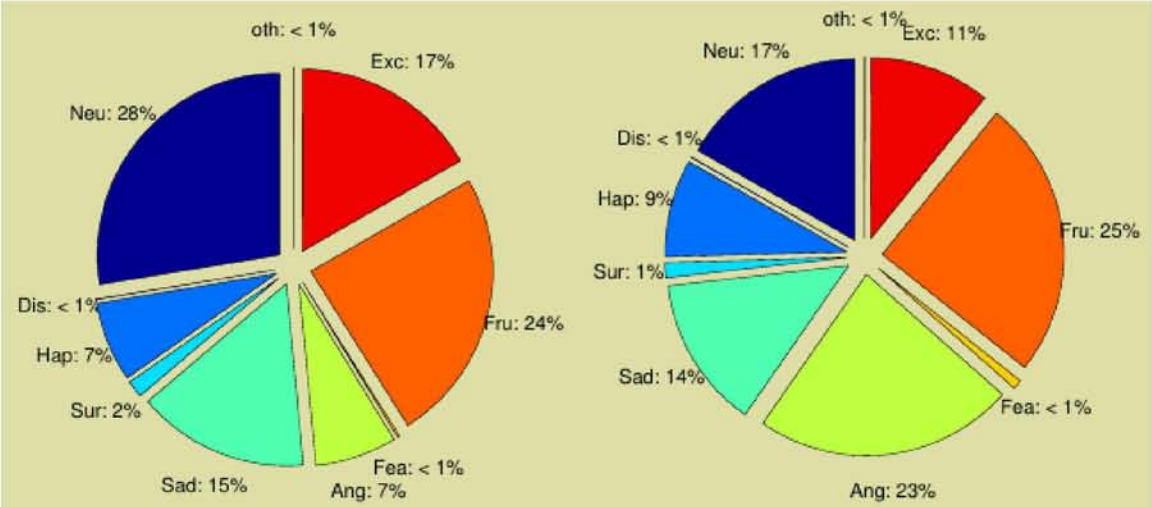
Σε αυτή την ενότητα θα μιλήσουμε για τις βάσεις δεδομένων που χρησιμοποιήθηκαν σε αυτή την εργασία. Στην επιλογή τους βοήθησαν οι [14] και [15], έρευνες που περιλαμβάνουν περιγραφές βάσεων που χρησιμοποιούνται στο πεδίο της αναγνώρισης συναισθημάτων. Η βάση δεδομένων που χρησιμοποιήθηκε για αυτή την έρευνα είναι η IEMOCAP: Interactive emotional dyadic motion capture database [16]. Είναι μια πλήρης συλλογή που περιλαμβάνει 12 ώρες οπτικοακουστικού υλικού, αλλά και δεδομένα για αναγνώριση χειρονομιών. Εφόσον μας ενδιαφέρει μόνο το κομμάτι του ήχου, δεν θα γίνει επέκταση στις άλλες δύο μορφές.

Για τις ηχογραφήσεις επιλέχθηκαν δέκα επαγγελματίες ηθοποιοί οι οποίοι κλήθηκαν να εκτελέσουν διάφορα σενάρια διαλόγων τα οποία χωρίστηκαν σε πέντε συνεδρίες. Σε κάθε μία από αυτές συμμετείχαν δύο ηθοποιοί, ένας άντρας και μία γυναίκα. Δόθηκαν τρία στημένα σενάρια διαλόγων κι επιπλέον τους ζητήθηκε να αυτοσχεδιάσουν πάνω σε υποθετικά σενάρια.

Στα πιο τεχνικά χαρακτηριστικά, οι ηχογραφήσεις περιλαμβάνουν ολόκληρους τους διαλόγους σε μορφή αρχείων wav με συχνότητα 16kHz. Τα αρχεία αυτά χωρίστηκαν περαιτέρω σε μικρότερα τμήματα ώστε κάθε ένα να περιέχει μόνο έναν ομιλητή. Με άλλα λόγια χώρισαν το διάλογο με βάση τη σειρά με την οποία μιλάει κάθε ηθοποιός. Αποτέλεσμα αυτού του διαχωρισμού είναι 10.039 (5255 από τα στημένα σενάρια και 4784 από τα αυθόρμητα σενάρια) αρχεία συναισθηματικών προτάσεων με μέση διάρκεια 4,5 δευτερολέπτων. Η ονομασία των αρχείων ακολουθεί ένα συγκεκριμένο πρότυπο. Το όνομα για κάθε αρχείο, για παράδειγμα, έχει τη μορφή Ses01F_impro01_M000_neu, σύμφωνα με το οποίο η πρόταση ανήκει στην πρώτη συνεδρία (Ses01F), ενώ οι ηθοποιοί εκτελούν το πρώτο σενάριο αυτοσχεδίασης (impro01) και αυτή είναι η πρώτη πρόταση για τον αρσενικό ηθοποιό (M000). Στο τέλος ακολουθεί και μια ετικέτα με το συναίσθημα που εκφράζεται μέσα στην πρόταση (neu). Το συναίσθημα του παραδείγματος είναι το ουδέτερο (neutral).

Η απεικόνιση των συναισθημάτων αποδίδεται και με τους δύο τρόπους που έχουμε αναφέρει. Για τις διακριτές κλάσεις χρησιμοποιήθηκαν οι ετικέτες θυμός, λύπη, χαρά, αναστάτωση, απέχθεια, φόβος, έκπληξη και η ουδέτερη κατάσταση (Εικόνα 2.2). Για την απόδοση ετικετών για το κάθε αρχείο, κλήθηκαν έξι αξιολογητές συνολικά. Κάθε αρχείο αξιολογήθηκε από τουλάχιστον τρεις για την μεγαλύτερη δυνατή αξιοπιστία. Στις περιπτώσεις όπου δεν μπορούσε να βγει σύμφωνη απόφαση η ετικέτα που σημειωνόταν ήταν το other.

Τέλος, η άλλη μέθοδος απεικόνισης περιλαμβάνει αξιολόγηση σε κάθε μία από τις τρεις διαστάσεις που χρησιμοποιούνται, σθένος, διέγερση, υπεροχή. Αριθμοί από το 1 έως το 5 σημειώνονται για κάθε μία διάσταση με το ένα να δηλώνει μειωμένη εμφάνιση του συγκεκριμένου χαρακτηριστικού και το 5 τη μέγιστη. Έτσι κάθε τμήμα έχει βαθμολογηθεί με τρεις επιμέρους αριθμούς, ένα για κάθε διάσταση.



Σχήμα 2.2: Κλάσεις συναισθημάτων της βάσης IEMOCAP [16].

2.2 Παρατηρήσεις

Κλείνοντας αυτό το κεφάλαιο θέλουμε να παραθέσουμε ορισμένες παρατηρήσεις-προβλήματα που μπορούν να προκύψουν κατά την αναζήτηση μιας βάσης δεδομένων.

- Διαθεσιμότητα των βάσεων. Αν και έχουν κατασκευαστεί πολλές, μόνο ένας μικρός αριθμός είναι διαθέσιμος στο ευρύ κοινό με όχι και τόσο μεγάλη ποσότητα δεδομένων (για χρήση νευρωνικών δικτύων), κάτι που είναι πολύ περιοριστικό για συστήματα με μεγαλύτερες φιλοδοξίες.

- Αποτέλεσμα της προηγούμενης παρατήρησης είναι η χρήση πολλών διαφορετικών συλλογών δεδομένων για την ίδια εργασία. Καθώς οι περισσότερες βάσεις διαφέρουν μεταξύ τους σε πολλούς τομείς, όπως τι συναισθήματα περιγράφουν, πως τα απεικονίζουν και σε τι γλώσσα είναι, προκύπτουν επιπλέον δυσκολίες για την επεξεργασία τους [17].

- Ποιότητα των ηχογραφήσεων. Αυτό οφείλεται στο τεχνικό κομμάτι, δηλαδή τα μέσα με τα οποία έγινε η ηχογράφηση, με αποτέλεσμα την κακή ποιότητα ήχου και της δυσκολίας ανάλυσής του. Επίσης οφείλεται και στην απόδοση των ομιλητών-

ηθοποιών. Τα συναισθήματα που εκφράζουν πολλές φορές είναι υπερβολικά στημένα με αποτέλεσμα οι αξιολογητές να μην έχουν καλό ποσοστό αναγνώρισης κάτι που αλλοιώνει την αξιοπιστία των δεδομένων.

- Βάσεις συγκεκριμένου σκοπού. Οι βάσεις δεδομένων κατά κύριο λόγο δημιουργούνται στα πλαίσια μιας έρευνας με πολύ συγκεκριμένους στόχους. Αυτό έχει ως αποτέλεσμα την δημιουργία συλλογών δεδομένων με πολύ περιορισμένες δυνατότητες όσον αφορά στην επαναχρησιμοποίησή τους για άλλες έρευνες.

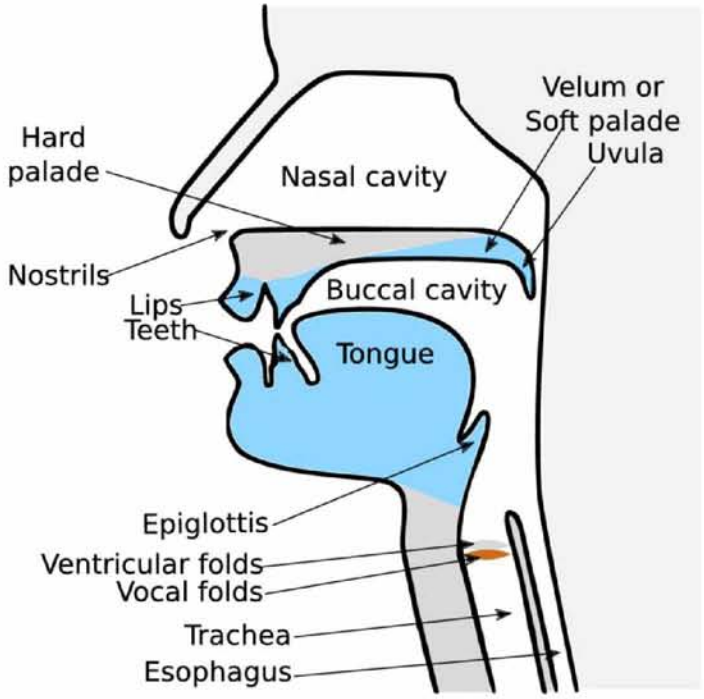
Κεφάλαιο 3

Εξαγωγή Χαρακτηριστικών

Στο Κεφάλαιο 2 αναφέραμε το πόσο σημαντική είναι η ύπαρξη κατάλληλων δεδομένων. Σε αυτό το κεφάλαιο θα αναλυθεί ο τρόπος με τον οποίο εξάγεται χρήσιμη πληροφορία, για τον εντοπισμό συναισθήματος, από τον ήχο. Η διαδικασία αυτή ονομάζεται εξαγωγή χαρακτηριστικών.

Ένα αρχείο wav δεν είναι παρά ένα σήμα ήχου και όπως όλα τα σήματα, μπορούν να επεξεργαστούν και να διαβαστούν από τους υπολογιστές. Η κύρια πηγή τέτοιων σημάτων είναι το σύστημα της φωνητικής οδού (Εικόνα 3.1), όπου οι δονήσεις των φωνητικών χορδών το διασχίζουν, διαμορφώνοντας έτσι τον τελικό ήχο που παράγεται. Σημαντικό ρόλο παίζει και η μορφολογία της φωνητικού συστήματος, η οποία διαφέρει από άνθρωπο σε άνθρωπο, και αλλάζει μορφή αναλόγως με το πόσο έντονα εκφράζεται το άτομο.

Τα ηχητικά σήματα, όμως, μεταβάλλονται με το χρόνο και μαζί τους η πληροφορία που περιέχουν καθιστώντας δύσκολη την ανάλυσή τους [19]. Ο τρόπος που αντιμετωπίζεται το πρόβλημα αυτό είναι η κατάτμηση του σήματος σε μικρές σχετικά περιοχές, τα πλαίσια, όπου σε κάθε ένα από αυτά η μεταβολή του χρόνου είναι αμελητέα, δίνοντας έτσι την αίσθηση μιας χρονικής αμεταβλητότητας. Η κατάτμηση, αλλά και μερικές ακόμη ενέργειες που θα αναλυθούν στη συνέχεια, αποτελούν στάδια της λεγόμενης προ-



Σχήμα 3.1: Φωνητικό σύστημα. Εικόνα από [18].

επεξεργασίας, όρος που χρησιμοποιείται για να δηλώσει την επεξεργασία του σήματος πριν την εξαγωγή οποιονδήποτε χαρακτηριστικών ([20], [21]).

1. **Φιλτράρισμα (Filtering):** Στόχος είναι η μείωση του θορύβου, που μπορεί να προέρχεται από τις συνθήκες ηχογράφησης με αποτέλεσμα την πιο αξιόπιστη συλλογή χαρακτηριστικών. Γίνεται χρήση φίλτρου υψηλής διέλευσης για την ενίσχυση της ενέργειας του σήματος στις υψηλές συχνότητες και την απόκτηση περισσότερων πληροφοριών από αυτές.

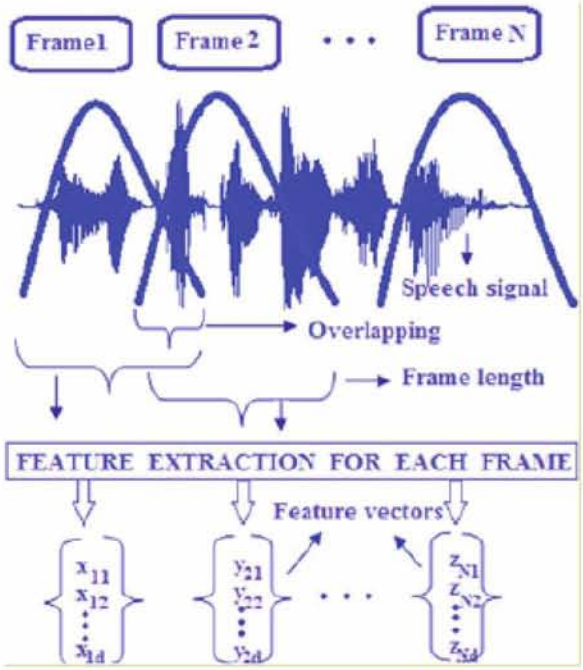
2. **Πλαισίωση (Framing):** Είναι η διαδικασία της κατάτμησης που αναφέρθηκε στην προηγούμενη παράγραφο. Το σήμα μετατρέπεται σε μια σειρά από πλαίσια ίδιου μεγέθους (μέσα σε ένα πλαίσιο υπάρχουν N δείγματα σήματος), με τα πιο συνηθισμένα μεγέθη πλαισίου να είναι 20-40ms. Το μέγεθός τους δεν θα πρέπει να είναι ούτε πολύ μεγάλο, ούτε πολύ μικρό. Αν το πλαίσιο είναι πολύ μεγάλο παρατηρούνται πολλές αλλαγές στο σήμα κάτι που απαλείφει την ιδιότητα της χρονικής αμεταβλητότητας. Από την άλλη πλευρά, αν το πλαίσιο είναι πολύ μικρό

δεν διαθέτει πολλά δείγματα από τα οποία θα προκύψουν τα χαρακτηριστικά. Από κάθε πλαίσιο εξάγονται ορισμένα χαρακτηριστικά που μας ενδιαφέρουν. Τα πλαίσια είναι επικαλυπτόμενα και προκύπτουν με μετατόπιση του αρχικού πλαισίου συνήθως ανά 10ms.

3. **Παραθύρωση (Windowing):** Κατά τον διαχωρισμό του σήματος σε πλαίσια παρουσιάζονται ασυνέχειες στα άκρα τους, οι οποίες δημιουργούν προβλήματα για κάποιους υπολογισμούς (π.χ ανάλυση Fourier). Για την αντιμετώπιση αυτού του προβήματος κάθε πλαίσιο πολλαπλασίαζεται με ένα παράθυρο. Το ιδανικό παράθυρο που χρησιμοποιείται τις περισσότερες φορές είναι το Hamming και περιγράφεται από την εξίσωση:

$$w[n] = \begin{cases} 0.54 - 0.46 \times \cos\left(\frac{2\pi n}{L-1}\right) & 0 \leq n \leq L-1 \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

Τα βήματα της προ-επεξεργασίας συνοψίζονται στην Εικόνα 3.2.



Σχήμα 3.2: Προ-επεξεργασία σήματος ([20]).

3.1 Κατηγορίες χαρακτηριστικών

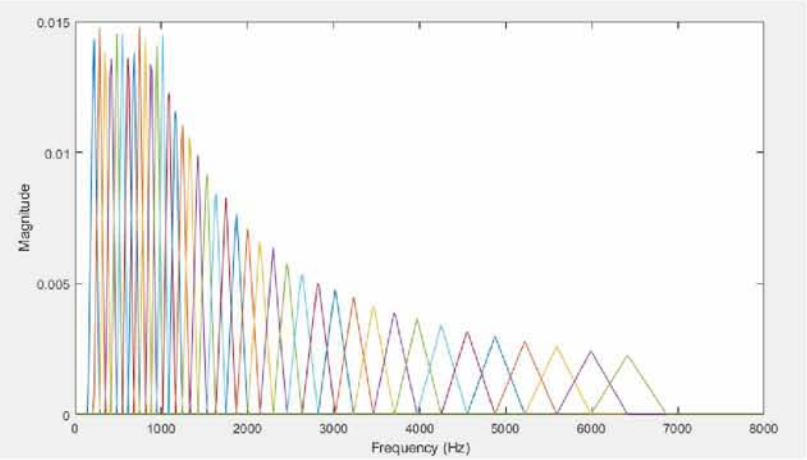
Τα χαρακτηριστικά που μπορούν να εξαχθούν από τα σήματα ορισμένες φορές κατηγοριοποιούνται ανάλογα με το μέσο που τα προκάλεσε. Για παράδειγμα, στο [22], τα χαρακτηριστικά που περιγράφουν τις δονήσεις των φωνητικών χορδών ανήκουν στην κατηγορία των πηγαίων χαρακτηριστικών, ενώ τα χαρακτηριστικά που περιγράφουν τη συμπεριφορά της φωνητικής οδού ονομάζονται χαρακτηριστικά συστήματος. Στις περισσότερες έρευνες όμως, κατατάσσονται σε χαρακτηριστικά φάσματος και προσωδιακά χαρακτηριστικά, τα οποία θα αναλυθούν παρακάτω.

Χαρακτηριστικά Φάσματος: Σε αυτή την κατηγορία επικεντρώνεται η εργασία αυτή. Τα φασματικά χαρακτηριστικά είναι από τα πιο διαδεδομένα σε έρευνες που σχετίζονται με αναγνώριση συναισθήματος και μετατροπή ομιλίας σε κείμενο ([20],[22],[23]). Είναι χαρακτηριστικά που προκύπτουν από το πεδίο συχνότητας του σήματος και προσομοιώνουν τη συμπεριφορά της φωνητικής οδού (χαρακτηριστικά συστήματος). Τα πιο γνωστά είναι:

- *Συντελεστές MFCC (Mel Frequency Cepstral Coefficients):* Δημιουργοί είναι οι Davis και Melstein στη δεκαετία του 1980 και μέχρι και σήμερα αποτελούν την πιο σύγχρονη μέθοδο. Ακολουθούν τα αναλυτικά βήματα για την εξαγωγή αυτών των χαρακτηριστικών, καθώς αυτά χρησιμοποιούνται από το σύστημα αναγνώρισης της έρευνας αυτής. Για κάθε πλαίσιο που έχει δημιουργηθεί από το στάδιο της προ-επεξεργασίας εφαρμόζονται τα εξής:

  1. **Βήμα 1:** Γίνεται μετατροπή στο φάσμα της συχνότητας με τη μέθοδο DFT (Discrete Fourier Transform) ή FFT (Fast Fourier Transform).
  2. **Βήμα 2:** Στο φάσμα που προκύπτει εφαρμόζεται μια σειρά φίλτρων, ή αλλιώς τράπεζα φίλτρων, που ονομάζονται Mel-filters. Αυτά τα φίλτρα σκοπό έχουν να μιμηθούν τον τρόπο με τον οποίο αντιλαμβάνονται τα αυτιά του ανθρώπου μεταβολές στη συχνότητα. Οι άνθρωποι αντιλαμβάνονται μικρές

αλλαγές στην ομιλία σε χαμηλές συχνότητες παρά σε υψηλές. Αυτός είναι ο λόγος που τα ζωνοπερατά αυτά φίλτρα είναι κατανεμημένα όπως φαίνεται στην Εικόνα 3.3.



Σχήμα 3.3: Mel-Filterbanks ([20]).

3. **Βήμα 3:** Υπολογίζεται ο λογάριθμος του φάσματος Mel.
4. **Βήμα 4:** Οι συντελεστές προκύπτουν από την εφαρμογή του DCT (Discrete Cosine Transform) στο λογαριθμημένο φάσμα Mel.

Τέλος, το πιο συνηθισμένο είναι να διαλέγονται οι 13 χαμηλότεροι συντελεστές οι οποίοι δημιουργούν το διάνυσμα χαρακτηριστικών που θα τροφοδοτήσει τους ταξινομητές. Αναλυτικά η διαδικασία εξαγωγής στην Εικόνα 3.4.

Σχήμα 3.4: Διαδικασία εξαγωγής MFCC.

- *Συντελεστές LPCC (linear Prediction Cepstral Coefficients).*
- *Συντελεστές PLP (Perceptual Linear Predictive).*

Προσωδιακά χαρακτηριστικά: Στην βιβλιογραφία τα συναντάμε και ως υπερκαλυπτικά χαρακτηριστικά (supra-segmental features) και είναι αυτά που εκφράζουν

τον ρυθμό, τον τόνο και τη μελωδία της φωνής κατά την εναλλαγή συναισθηματικών καταστάσεων ([9], [22]). Η ονομασία "υπερκαλυπτικά" προκύπτει από το γεγονός ότι εξάγονται από μεγαλύτερα τμήματα από αυτά των πλαισίων, όπως για παράδειγμα από ολόκληρη πρόταση. Σύμφωνα με το [20], αυτό το είδος χαρακτηριστικών μπορεί να διαχωρίζει αποτελεσματικότερα συναισθήματα μεγάλης διέγερσης με αυτά που έχουν λιγότερη.

Τα τρία βασικά χαρακτηριστικά είναι:

- *Zero Crossing Rate:* Μετράει πόσες φορές το πλάτος του σήματος θα περάσει από τη μηδενική τιμή μέσα σε ένα διάστημα (π.χ πλαίσιο). Εάν η συχνότητα των διασχίσεων της μηδενικής τιμής είναι υψηλή σημαίνει πως το σήμα μάλλον είναι άφωνο, ενώ αν το η συχνότητα είναι χαμηλή, το σήμα είναι έμφωνο .

- *Short time Energy:* Είναι η μέτρηση του πλάτους του σήματος το οποίο μεταβάλλεται με την πάροδο του χρόνου. Καταγράφει τις αλλαγές αυτές, και αν η τιμή του χαρακτηριστικού αυτού είναι μεγάλη, σημαίνει πως το σήμα ήχου περιέχει φωνή, αντιθέτως στην περίπτωση χαμηλής τιμής της ενέργειας του σήματος έχουμε έλλειψη ομιλίας.

- *Ένταση - Pitch*: Εκφράζει τη θεμελιώδη συχνότητα με την οποία πάλλονται οι φωνητικές χορδές, η οποία αυξομειώνεται κατά τη διάρκεια της ομιλίας.

3.2 Εργαλεία εξαγωγής χαρακτηριστικών

Υπάρχουν πολλά εξελιγμένα λογισμικά τα οποία μπορούν να εξάγουν μια πληθώρα χαρακτηριστικών και να εφαρμόσουν τεχνικές προ-επεξεργασίας. Μερικά από αυτά είναι το Kaldi, openSMILE, Praat. Στην εργασία αυτή χρησιμοποιήθηκε το εργαλείο Kaldi, το οποίο αναλύεται στο Κεφάλαιο 4.

Το λογισμικό αυτό έχει πολλές δυνατότητες εξαγωγής χαρακτηριστικών. Μπορεί να υπολογίσει MFCC συντελεστές και υπάρχει δυνατότητα για παραμετροποίηση της διαδι-

κασίας αυτής. Για παράδειγμα, μέσα από ένα αρχείο διαμόρφωσης μπορούν να οριστούν οι τιμές για το μέγεθος των πλαισίων ή το είδος του παραθύρου που θα χρησιμοποιηθεί. Επίσης, χρησιμοποιεί αλγόριθμους για τη μείωση των διαστάσεων του διανύσματος των χαρακτηριστικών, όπως και μεθόδους προ-επεξεργασίας και ομαλοποίησης των δεδομένων.

Βήματα εξαγωγής συντελεστών MFCC από το Kaldi:

- Αρχικά χρησιμοποιεί πλαίσια των 25ms με μετατόπιση κατά 10ms και για κάθε πλαίσιο,
- εξάγει τα δεδομένα και τα πολλαπλασιάζει με ένα παράθυρο. Στην περίπτωση της εργασίας αυτής επιλέχθηκε μια παραλλαγή του Hamming ειδικά φτιαγμένη από τους δημιουργούς του Kaldi.
- Εφαρμόζεται ο αλγόριθμος FFT για τη μετατροπή στο πεδίο της συχνότητας.
- Γίνεται ο υπολογισμός των τιμών του φάσματος Mel.
- Υπολογισμός του λογάριθμου των ενεργειών και εφαρμογή του DCT από τον οποίο προκύπτουν 13 συντελεστές MFCC.

Στην αναγνώριση ομιλίας είναι σύνηθες να υπολογίζονται επιπλέον συντελεστές, οι οποίοι προκύπτουν από τους αρχικούς και ονομάζονται διαφορικοί συντελεστές Delta, αλλά και συντελεστές επιτάχυνσης Delta-Deltas, που προκύπτουν από τους διαφορικούς.

Κεφάλαιο 4

Μέθοδοι και εργαλεία ταξινόμησης

4.1 Μοντέλα Ταξινόμησης

Ο τρόπος με τον οποίο μαθαίνουν να αναγνωρίζουν πρότυπα τα μοντέλα ταξινόμησης αυτής της εργασίας, γίνεται με την τεχνική της **μάθησης με επίβλεψη**. Σύμφωνα με αυτή, ένας ταξινομητής τροφοδοτείται με δεδομένα από τα οποία μαθαίνει να αναγνωρίζει ορισμένα χαρακτηριστικά. Τα δεδομένα διαθέτουν μία ετικέτα και ο ταξινομητής αντιστοιχίζει την ετικέτα με τα χαρακτηριστικά που διάβασε. Αφού αποκτήσει μια ικανή γνώση για αυτά, προσπαθεί στη συνέχεια να εντοπίσει παρόμοια χαρακτηριστικά σε δεδομένα τα οποία δεν έχει συναντήσει και να εξάγει μία ετικέτα για να τα κατηγοριοποιήσει.

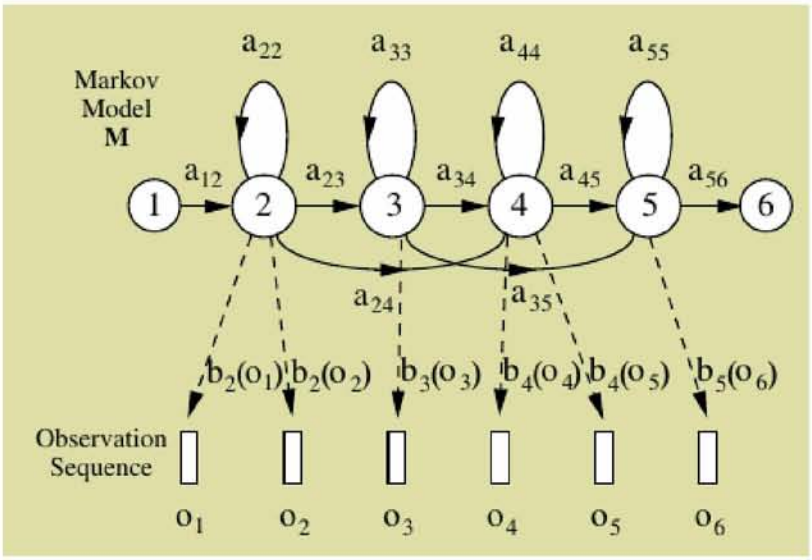
Εν συντομία, το βασικό μοντέλο ταξινομητή που χρησιμοποιείται είναι τα Κρυφά Μοντέλα Markov. Ο ρόλος των νευρωνικών δικτύων αναδεικνύεται στην μοντελοποίηση των κατανομών πιθανοτήτων των μοντέλων Markov. Πριν όμως από τα νευρωνικά δίκτυα, γίνεται μία προσέγγιση των κατανομών αυτών και με Γκαουσιανά Μοντέλα Μίξης (Gaussian Mixture Models).

4.1.1 Κρυφά Μοντέλα Markov (HMM)

Τα HMMs είναι στοχαστικά μοντέλα Markov τα οποία μοντελοποιούν τυχαίως μεταβαλλόμενα συστήματα με δεδομένα που έχουν υπόσταση στον χρόνο (καιρός, κείμενα, ομιλία). Μέσα από μια ακολουθία παρατηρήσεων-δεδομένων, μπορούν να εξάγουν κάποιο στατιστικό συμπέρασμα σχετικά με την κατάσταση στην οποία βρίσκεται ένα σύστημα.

Δομή των κρυφών μοντέλων Markov

Η δομή τους είναι ίδια με αυτή των μοντέλων Markov. Είναι στοχαστικά πεπερασμένα αυτόματα και διαθέτουν ένα σύνολο καταστάσεων, έναν πίνακα πιθανότητας μεταβάσεων, και αρχικές πιθανότητες κατάστασης. Η διαφορά τους όμως με τα μοντέλα Markov είναι ότι οι καταστάσεις είναι κρυμμένες από τον παρατηρητή και το μόνο που φαίνεται είναι μια αλληλουχία παρατηρήσεων. Η δουλειά των HMMs είναι να κάνουν μια εκτίμηση για τη σειρά των καταστάσεων, έχοντας ως δεδομένο μια αλληλουχία παρατηρήσεων. Η αναλυτική τοπολογία τους παρουσιάζεται στο σχήμα 4.1.



Σχήμα 4.1: Τοπολογία Κρυφού Μοντέλου Markov [24].

Ένα HMM ορίζεται από τις παραμέτρους $\lambda = \{A, B, \pi\}$ [25]:

- Σύνολο N καταστάσεων $\mathbf{S} = \{s_1, s_2, ..., s_N\}$, όπου το s_t είναι η κατάσταση στην χρονική στιγμή t.
- Πίνακα πιθανοτήτων μετάβασης $\mathbf{A} = \{a_{ij}\}$, όπου a_{ij} είναι η πιθανότητα να γίνει η μετάβαση από την κατάσταση i στην κατάσταση j.
- Σύμβολα παρατηρήσεων $\mathbf{O} = \{o_1, o_2, ..., o_M\}$.
- Κατανομή πιθανοτήτων παρατήρησης $\mathbf{B} = \{b_i(o_t)\}$, όπου $b_i(o_t)$ είναι η πιθανότητα να έχουμε ως έξοδο το σύμβολο o_t όταν βρισκόμαστε στην κατάσταση i.
- Κατανομή αρχικών καταστάσεων $\pi = \{\pi_i\}$.

Έχοντας διευκρινίσει τις παραμέτρους, ορίζουμε ένα HMM ως μια μηχανή πεπερασμένων καταστάσεων που αλλάζει καταστάσεις ανά μονάδα χρόνου και κάθε χρονική στιγμή t, όπου το σύστημα μεταβαίνει στην κατάσταση s_t, παράγεται μια παρατήρηση (π.χ ένα διάνυσμα χαρακτηριστικών) από την πυκνότητα πιθανότητας $b_i(k)$. Επιπλέον, μια μελλοντική κατάσταση/γεγονός εξαρτάται μόνο από την τρέχουσα κατάσταση/γεγονός και όχι από κάποια παλαιότερη, δηλαδή $p(s_t|s_{t-1})$ (ιδιότητα Markov).

Η χρήση μοντέλων Markov επιβάλλει την επίλυση τριών βασικών προβλημάτων [25]:

1. **Πρόβλημα εκτίμησης:** Ο προσδιορισμός της συνολικής πιθανότητας, ώστε μια αλληλουχία παρατηρήσεων $\{o_1, o_2, ..., o_T\}$ να προέρχεται από ένα HMM.
2. **Πρόβλημα αποκωδικοποίησης:** Ο προσδιορισμός της πιο πιθανής αλληλουχίας κρυφών καταστάσεων, έχοντας ως δεδομένο μια αλληλουχία παρατηρήσεων κι ένα HMM.
3. **Πρόβλημα εκπαίδευσης:** Ο προσδιορισμός των βέλτιστων παραμέτρων $\mathbf{A}$ και $\mathbf{B}$ για μια αλληλουχία παρατηρήσεων.

Για την επίλυση του πρώτου προβλήματος χρησιμοποιείται ο αλγόριθμος Forward. Για να υπολογίσουμε την πιθανότητα παραγωγής μιας σειράς παρατηρήσεων, δεδομένου ενός

HMM, αθροίζουμε τις πιθανότητες των παρατηρήσεων για όλες τις πιθανές αλληλουχίες καταστάσεων, δηλαδή υπολογίζεται το

$$p(O) = \sum_S p(O, S) = \sum_S p(O|S)p(S). \tag{4.1}$$

Ο υπολογισμός αυτός δεν είναι αποδοτικός για αυτό χρησιμοποιείται ο αλγόριθμος Forward, ο οποίος ορίζει την εμπρόσθια πιθανότητα (forward probability), με αναδρομικό τρόπο, ως εξής:

$$a_t(s_j) = \left[\sum_{i=1}^{N} a_{t-1}(s_i)a_{ij}\right] b_j(o_t). \tag{4.2}$$

Η $a_t(s_j)$ είναι η πιθανότητα να βρισκόμαστε στην κατάσταση j τη χρονική στιγμή t και να έχουμε παρατηρήσει τις πρώτες t παρατηρήσεις $o_1, o_2, ..., o_t$. Η εξίσωση 4.2 ορίζει πως η πιθανότητα $a_t(s_j)$ υπολογίζεται αθροίζοντας όλες τις εμπρόσθιες πιθανότητες όλων των προηγούμενων καταστάσεων με συντελεστή βάρους τις πιθανότητες μετάβασης a_{ij}. Τα βήματα του αλγόριθμου είναι τα εξής:

- Αρχικοποίηση:

$$a_1(s_j) = \pi_j b_j(o_1), \quad 1 \leq j \leq N$$

- Αναδρομή: Για κάθε χρονική στιγμή,

$$a_t(s_j) = \left[\sum_{i=1}^{N} a_{t-1}(s_i)a_{ij}\right] b_j(o_t), \quad 1 \leq j \leq N, \quad 1 < t \leq T \tag{4.3}$$

- Τερματισμός: Η τελική λύση είναι

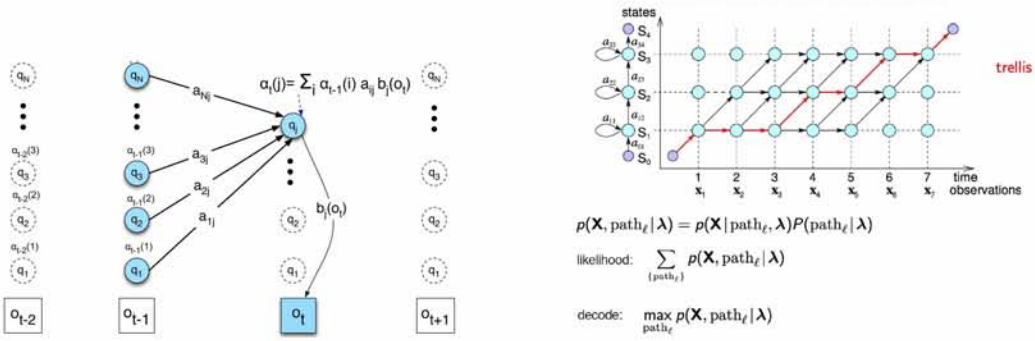
$$p(O|\lambda) = a_T(s_{final}) = \sum_{t=1}^{N} a_T(s_i)a_{ifinal}. \tag{4.4}$$

Ένα παράδειγμα υλοποίησης φαίνεται και στο γράφημα trellis της Εικόνας 4.2α΄, όπου ο οριζόντιος άξονας αντιπροσωπεύει τον χρόνο και ο κάθετος τις καταστάσεις του HMM.

Με παρόμοιο τρόπο ορίζεται και άλλο ένα βασικό σύνολο πιθανοτήτων, οι backward probabilities, οι οποίες εκφράζουν την πιθανότητα των μελλοντικών παρατηρήσεων από τη χρονική στιγμή $t+1$ μέχρι το τέλος, δεδομένου του μοντέλου HMM και της κατάστασης τη στιγμή t:

$$\beta_t(s_j) = p(o_{t+1}, ..., o_T | S(t) = s_j, \lambda) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(s_j). \tag{4.5}$$

Για το πρόβλημα της αποκωδικοποίησης χρησιμοποιείται μια παρόμοια μέθοδος με την προηγούμενη, ο αλγόριθμος Viterbi, ο οποίος βρίσκει την καλύτερη αλληλουχία καταστάσεων βάσει των παρατηρήσεων και ενός μοντέλου HMM. Η μόνη διαφορά είναι ότι αντί του αθροίσματος στην εξίσωση 4.4, χρησιμοποιείται η συνάρτηση μεγιστοποίησης, αλλά και οπισθοδείκτες (backpointers) που καταγράφουν το μονοπάτι των πιθανότερων καταστάσεων. Οπτικά, ψάχνει να βρει το καλύτερο μονοπάτι σε ένα γράφημα trellis. (Εικόνα 4.2β′).



(α′) Forward trellis (β′) Viterbi trellis

Σχήμα 4.2: Γραφήματα trellis [26]

Το τελευταίο και πιο πολύπλοκο πρόβλημα είναι ο υπολογισμός των παραμέτρων ενός HMM. Ο αλγόριθμος Baum–Welch (forward-backward algorithm) αποτελεί ειδική κατηγορία του αλγόριθμου Αναμενόμενης Τιμής - Μεγιστοποίησης (Expectation-Maximization (EM)) και συνοψίζεται στα εξής βήματα. Πρώτα αρχικοποιούνται οι τιμές των παραμέτρων του HMM με τυχαίο τρόπο σε περίπτωση που δεν υπάρχει πληροφορία για αυτές. Στη συνέχεια, εφαρμόζεται το στάδιο E (Expectation) κατά το οποίο υπολο-

γίζονται αναδρομικά οι πιθανότητες $a_t(s_j)$ και $\beta_t(s_j)$ που αναφέρθηκαν παραπάνω. Το κύριο σημείο είναι ο υπολογισμός των πιθανοτήτων κατοχής κατάστασης και διέλευσης. Έτσι αντίστοιχα, με $\gamma_t(s_j)$ συμβολίζεται η πιθανότητα να βρισκόμαστε στην κατάσταση j τη χρονική στιγμή t, ενώ με $\xi_t(i,j)$ συμβολίζεται η πιθανότητα να βρισκόμαστε στην κατάσταση i τη χρονική στιγμή t και στην κατάσταση j τη χρονική στιγμή tt+1. Όλα αυτά έχοντας ως δεδομένο μια αλληλουχία παρατηρήσεων. Οι εξισώσεις που περιγράφουν τις πιθανότητες αυτές είναι:

$$\gamma_t(s_j) = p(s_{t=j}|O,\lambda) = \frac{1}{\alpha_T(s_{final})}\alpha_t(s_j)\beta_t(s_j), \tag{4.6}$$

και

$$\xi_t(i,j) = p(s_{t=i}, s_{t+1=j}|O,\lambda) = \frac{\alpha_t(s_i)a_{ij}b_j(o_{t+1})\beta_{t+1}(s_j)}{\alpha_T(s_{final})}. \tag{4.7}$$

Οι πιθανότητες αυτές στην ουσία υπολογίζουν για κάθε κατάσταση την πιθανότητα να έχουν παρατηρήσει ένα γεγονός σε αντίστοιχη χρονική στιγμή. Με τον συνυπολογισμό τους, για όλους τους συνδυασμούς καταστάσεων, προκύπτει μια ευθυγράμμιση μεταξύ των παρατηρήσεων και των καταστάσεων του μοντέλου. Η ευθυγράμμιση αυτή είναι πιο αντιπροσωπευτική από την αρχική η οποία έγινε με τυχαίο τρόπο και αυτή είναι η λειτουργία αυτού του τμήματος του αλγόριθμου.

Το επόμενο βήμα είναι αυτό της μεγιστοποίησης M (Maximization). Εφόσον εξήχθη μια εκτίμηση για τις κατανομές των $\gamma_t(s_j)$ και $\xi_t(i,j)$, γίνεται επανεκτίμηση των παραμέτρων του HMM. Τα δύο αυτά βήματα (E-M) εκτελούνται επαναληπτικά μέχρι τα αποτελέσματα να συγκλίνουν ή να έχει ορίσει ο χρήστης ένα συγκεκριμένο αριθμό επαναλήψεων της διαδικασίας. Η γενική φιλοσοφία του αλγόριθμου είναι ότι διορθώνοντας ένα σύνολο παραμέτρων, βελτιώνεται κάποιο άλλο και αυτή η διαδικασία συνεχίζεται έως ότου να συγκλίνουμε σε μία λύση.

Εναλλακτικά, για τον προσδιορισμό των παραμέτρων των HMMs χρησιμοποιείται και ο αλγόριθμος Viterbi για εκπαίδευση. Η σημαντική διαφορά σε σχέση με τον αλγόριθμο

Baum-Welch είναι, ότι επιλέγει την ευθυγράμμιση με την μεγαλύτερη πιθανότητα χωρίς να συνυπολογίζει τους υπόλοιπους συνδυασμούς. Κι εδώ γίνεται επανεκτίμηση των παραμέτρων του μοντέλου και η διαδικασία αυτή γίνεται επαναληπτικά 4.3. Οι παραπάνω

(α′) Πρόβλημα αποκωδικοποίησης (β′) Πρόβλημα εκτίμησης παραμέτρων

Σχήμα 4.3: Στάδια εκπαίδευσης αλγόριθμου Viterbi

αλγόριθμοι αναλύονται λεπτομερώς στο [25], αλλά και στο [24] όπου εξετάζονται από τη σκοπιά μοντέλων αναγνώρισης ομιλίας. Επίσης, σημαντικό είναι να αναφερθεί πως οι τιμές με τις οποίες γίνονται οι υπολογισμοί των αναδρομών στους παραπάνω αλγόριθμους, εκφράζονται λογαριθμικά. Αυτό συμβαίνει διότι οι υπολογισμοί περιέχουν πολλούς πολλαπλασιασμούς μεταξύ πιθανοτήτων (αριθμοί μικρότεροι της μονάδας) και υπάρχει ο κίνδυνος της υποχείλισης (underflow).

4.1.2 Γκαουσιανά Μοντέλα Μίξης

Ένα Γκαουσιανό Μοντέλο Μίξης GMM είναι μια κατανομή πιθανοτήτων. Όπως υποδηλώνει και το όνομα, αποτελείται από πολλαπλές Γκαουσιανές κατανομές. Ένα GMM μπορεί να γραφτεί ως:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k), \tag{4.8}$$

όπου το x ένα διάνυσμα d διαστάσεων, π_k το βάρος του k-οστού Γκαουσιανού στοιχείου, μ_k το d-διάστατο διάνυσμα μέσων τιμών για το k-οστό Γκαουσιανό στοιχείο και Σ_k είναι ο $d \times d$ πίνακας συνδιασποράς για το k-οστό Γκαουσιανό στοιχείο. Η συνάρτη-

ση $\mathcal{N}(x|\mu, \Sigma)$ είναι ο τύπος της Γκαουσιανής κατανομής ή διαφορετικά της κατανομής πυκνότητας πιθανότητας:

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}} \frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}. \tag{4.9}$$

Με τη χρήση του αλγόριθμου (EM), ο οποίος είναι μία προσέγγιση μέγιστης πιθανότητας, είναι δυνατόν να βρεθούν οι βέλτιστες παράμετροι του GMM ώστε να ταξινομήσει όσο καλύτερα μπορεί τα δείγματα.

4.1.3 Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα είναι ένα είδος ταξινομητών, που την τελευταία δεκαετία είναι πολύ δημοφιλή με μεγάλη αύξηση δημοσιευμένων ερευνών σε σχέση με παλαιότερα. Η έννοια των νευρωνικών δικτύων δεν είναι καινούργια, ο Frank Rosenblatt [27] εισήγαγε την έννοια του perceptron, της πιο απλής μορφής ενός νευρωνικού δικτύου την δεκαετία του '60. Η εξέλιξη, όμως, της τεχνολογίας και συγκεκριμένα η αύξηση της υπολογιστικής ισχύος στις σύγχρονες κάρτες γραφικών έδωσαν νέα ώθηση στην έρευνα του κλάδου αυτού. Ακολουθούν σύντομες περιγραφές ορισμένων δομών νευρωνικών δικτύων.

Αρχιτεκτονικές Νευρωνικών Δικτύων

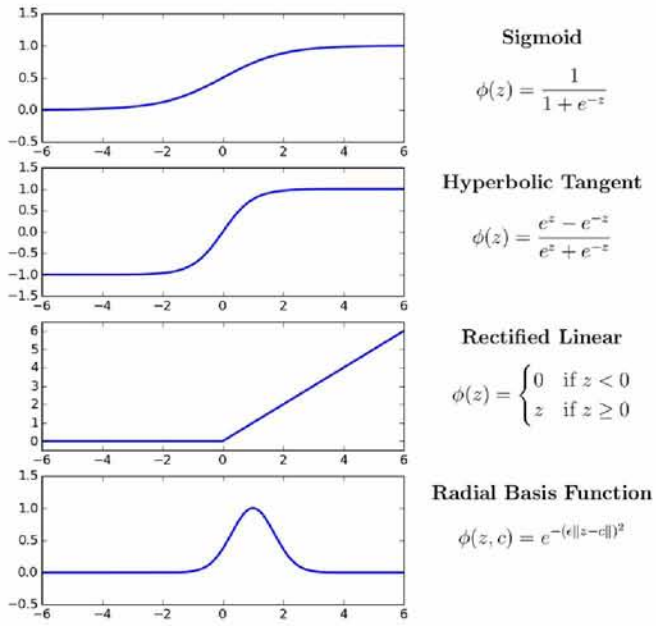
Η δομή και η φιλοσοφία γύρω από τα Τεχνητά Νευρωνικά Δίκτυα στηρίχτηκε στο βιολογικό μας δίκτυο, τον εγκέφαλο και τους νευρώνες από τους οποίους αποτελείται. Όπως οι νευρώνες είναι συνδεδεμένοι μεταξύ τους και ανταλλάσσουν σήματα, έτσι και τα Τεχνητά Νευρωνικά Δίκτυα αποτελούνται από ένα συνδεδεμένο πλέγμα κόμβων με τη δυνατότητα ανταλλαγής πληροφοριών.

Ένας κόμβος ονομάζεται σιγμοειδές perceptron (Εικόνα 4.5α') και αποτελείται από εισόδους με εύρος τιμών από 0 έως 1. Σε κάθε είσοδο συνυπολογίζεται επιπλέον, μια τιμή βάρους και μια σταθερή τιμή bias. Η έξοδός του ορίζεται από μία συνάρτηση που

ονομάζεται σιγμοειδής, παραδείγματα της οποίας φαίνονται στην εικόνα 4.4. Για εισόδους x_1, x_2 , ..., βάρη w_1, w_2 , ..., και bias b, η έξοδός του είναι:

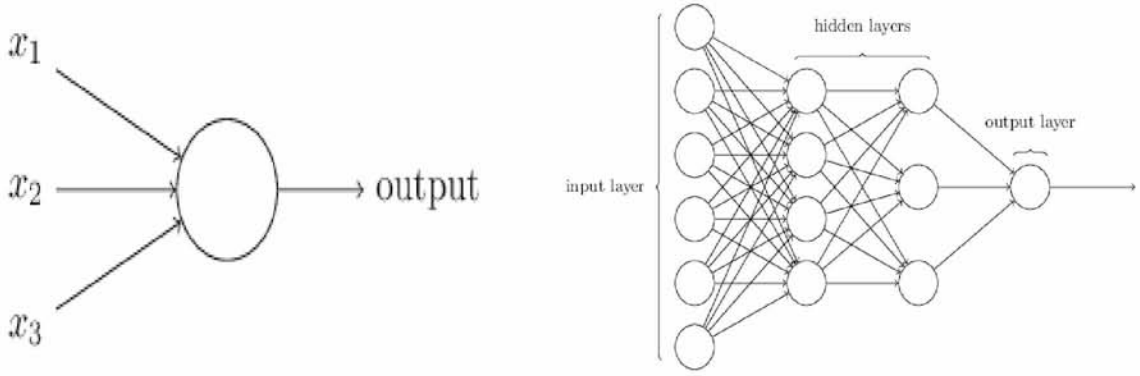
$$h_\theta(x) = \frac{1}{1 + e^{-\Sigma_j w_j x_j - b}} \tag{4.10}$$

Με αυτό το μηχανισμό, μια μικρή αλλαγή της εισόδου μπορεί να επιφέρει μια μικρή αλλαγή στην έξοδο ανάλογα με τις τιμές των βαρών, κάτι που σημαίνει ότι υπάρχει η δυνατότητα να εκπαιδευτεί το σύστημα ώστε να λύνει προβλήματα.



Σχήμα 4.4: Συναρτήσεις Ενεργοποίησης [28].

Όπως υποδηλώνεται επομένως, ένα Τεχνητό Νευρωνικό Δίκτυο αποτελείται από πολλά συνδεδεμένα μεταξύ τους perceptrons τα οποία είναι χωρισμένα σε επίπεδα. Το πρώτο και τελευταίο επίπεδο ονομάζονται επίπεδο εισόδου και επίπεδο εξόδου αντίστοιχα, ενώ τα ενδιάμεσα επίπεδα, αν υπάρχουν, λέγονται κρυφά επίπεδα 4.5β΄. Στην περίπτωση της αναγνώρισης προτύπων, οι είσοδοι είναι τα εξαγόμενα χαρακτηριστικά τα οποία περνούν μέσα από τα κρυφά επίπεδα και στην έξοδο ενεργοποιείται κάποιος συγκεκριμένος κόμβος-νευρώνας που αντιστοιχεί σε μια ετικέτα.



(α′) Perceptron (β′) Τεχνητό Νευρωνικό Δίκτυο

Σχήμα 4.5: Δομή Νευρωνικών Δικτύων [29]

Είδαμε πως η δομή τέτοιων δικτύων μπορεί να περιλαμβάνει από ένα έως περισσότερα κρυφά επίπεδοτα. Τα δίκτυα με παραπάνω από ένα επίπεδο νευρώνων-κόμβων, ονομάζονται Βαθιά Νευρωνικά Δίκτυα (DNNs) και είναι οι δομές που εξερευνούμε στα πειράματα του Κεφαλαίου 5. Μερικές από τις πιο γνωστές δομές, επιγραμματικά είναι τα Νευρωνικά Δίκτυα Συνέλιξης (CNN), τα Νευρωνικά Δίκτυα Ανατροφοδότησης (RNN) και πολλές παραλλαγές τους.

Για την εκπαίδευση των DNNs, ο πιο διαδεδομένος τρόπος είναι η μέθοδος gradient descent μαζί με την μέθοδο οπισθοδιάδοσης (backpropagation). Η βασική ιδέα πίσω από αυτά είναι ότι θέλουμε να βρούμε τα βάρη και τα biases, ώστε η έξοδος του δικτύου να προσεγγίζει την επιθυμητή τιμή για όλα τα δεδομένα εισόδου. Το πόσο καλή είναι αυτή η προσέγγιση ελέγχεται από μία συνάρτηση κόστους. Μια χαρακτηριστική είναι η συνάρτηση τετραγωνικού σφάλματος (squared error function):

$$C(w,b) = \frac{1}{2n}\Sigma_x(y(x) - a)^2, \tag{4.11}$$

με w να είναι όλα τα βάρη, b όλα τα biases και a το διάνυσμα των εξόδων. Το άθροισμα γίνεται για όλα τα δεδομένα εισόδου. Ο αλγόριθμος gradient descent βρίσκει για ποιες τιμές ελαχιστοποιείται η συνάρτηση κόστους. Είναι μια επαναληπτική μέθοδος

βελτιστοποίησης όπου σε κάθε επανάληψη έχουμε:

$$\theta_{j+1} = \theta_j - \alpha \frac{\partial}{d\theta_j} J(\theta_0, \theta_1), \tag{4.12}$$

με $J(\theta_0, \theta_1)$ είναι η συνάρτηση κόστους, τα θ_0, θ_1 συμβολίζουν το bias και τα βάρη αντίστοιχα, και το α είναι ο ρυθμός μάθησης, παράμετρος που επηρεάζει την ταχύτητα και την αποτελεσματικότητα του αλγόριθμου. Τέλος, η συμβολή της οπισθοδιάδοσης έγκειται στον αποδοτικό υπολογισμό κλίσεων (gradient).

4.1.4 Μοντέλα GMM-HMM και DNN-HMM

GMM-HMM

Όπως αναφέρθηκε, η εργασία αυτή χρησιμοποιεί τα μοντέλα GMM-HMM τα οποία είναι HMMs με τις πιθανότητες εκπομπής τους να μοντελοποιούνται από ένα γραμμικό συνδυασμό Γκαουσιανών κατανομών. Αυτό έχει ως αποτέλεσμα οι παράμετροι εξόδου, $b_i(k)$, να εκφράζονται σύμφωνα με τις παραμέτρους του μοντέλου μίξης (μ, Σ), δηλαδή

$$b_j(o_t) = p(o_t|S = s_j) = \sum_{k=1}^{K} c_{jk} \mathcal{N}(o_t|\mu_{jk}, \Sigma_{jk}), \tag{4.13}$$

όπου η παράμετρος c_{jk} είναι ο συντελεστής-βάρος της k-οστής Γκαουσιανής κατανομής από τις οποίες αποτελείται το μοντέλο μίξης. Τα GMMs παρουσιάζουν τη σχέση μεταξύ των παρατηρήσεων και των κρυφών καταστάσεων. Ως προς την εκπαίδευση των μοντέλων, εξακολουθούν να ισχύουν οι ίδιοι αλγόριθμοι που συζητήθηκαν στην ενότητα 4.1.1, προσαρμοσμένοι βέβαια στις νέες παραμέτρους [25]. Οι διαστάσεις των Γκαουσιανών εξαρτώνται από τις διαστάσεις των παρατηρήσεων.

DNN-HMM

Σε περιπτώσεις που τα δεδομένα παρουσιάζουν μη-γραμμικές σχέσεις, τα βαθιά νευρωνικά δίκτυα είναι αποδοτικότερα, καθώς τα μοντέλα μίξης δεν μπορούν να μοντελοποιήσουν τέτοιες ιδιότητες. Για αυτό το λόγο, μοντέλα DNN-HMM χρησιμοποιούνται για την επίλυση προβλημάτων αναγνώρισης ήχου, χειρονομιών και συναισθημάτων [30]. Η διαφορά τους από τα μοντέλα GMM-HMM, είναι ότι η εκτίμηση των πιθανοτήτων των παρατηρήσεων γίνεται από βαθιά νευρωνικά δίκτυα και όχι από Γκαουσιανά μοντέλα μίξης [31]. Ωστόσο, όπως θα περιγραφεί και στη συνέχεια του κεφαλαίου, η συμβολή των μοντέλων GMM-HMM είναι απαραίτητη στην διαδικασία εκπαίδευσης τέτοιων μοντέλων.

Προτού αναφερθούν οι λεπτομέρειες υλοποίησης του συστήματος αυτής της εργασίας, η επόμενη υποενότητα τονίζει την χρησιμότητα των μοντέλων Markov και ποια προβλήματα είναι ικανά να επιλύσουν. Η χρήση τους HMMs παρατηρείται κυρίως σε εφαρμογές αυτόματης αναγνώρισης ομιλίας, καθώς επίσης και για την αναγνώριση χειρονομιών. Υπάρχουν, όμως, και έρευνες που εξετάζουν υλοποιήσεις για την αναγνώριση συναισθημάτων [32], όπως και αυτή η εργασία.

4.2 Χρήση HMM στην αναγνώριση ομιλίας και συναισθήματος

Ο τρόπος που υλοποιούνται τα μοντέλα αναγνώρισης συναισθήματος σε αυτή την εργασία στηρίζεται στους μηχανισμούς της αυτόματης αναγνώρισης ομιλίας. Σε αυτή την ενότητα περιγράφεται, με μια γενική εικόνα, η βασική ιδέα πίσω από τα συστήματα αναγνώρισης ομιλίας και πως αυτή προσαρμόζεται για την αναγνώριση συναισθημάτων. Μια αναλυτικότερη σύγκριση των δύο αυτών συστημάτων παρουσιάζεται στο [33].

Αναγνώριση Ομιλίας

Στόχος ενός τέτοιου συστήματος είναι η επίλυση του πιθανοτικού προβλήματος εύρεσης της πιο πιθανής αλληλουχίας λέξεων **W**, έχοντας ως δεδομένο μια σειρά ακουστικών παρατηρήσεων **O**:

$$W = \underset{W}{\operatorname{argmax}}\, P(W|O) \quad \overset{\text{Bayes}}{\Longrightarrow} \quad W = \underset{W}{\operatorname{argmax}}\, P(O|W)P(W),$$

όπου το $P(W|O)$ αντιπροσωπεύει το ακουστικό μοντέλο, ενώ το $P(W)$ εκφράζει το γλωσσικό μοντέλο.

Η μοντελοποίηση του ακουστικού μοντέλου γίνεται από HMMs τα οποία μπορεί να συνοδεύονται από Γκαουσιανά μοντέλα μίξης ή νευρωνικά δίκτυα. Η ιδέα είναι ότι κάθε φώνημα αντιπροσωπεύεται από ένα τέτοιο μοντέλο και η αναγνώριση προκύπτει από μια αναζήτηση Viterbi στο σύνολο όλων των μοντέλων, από όπου προκύπτουν οι λέξεις που αναγνωρίζονται.

Σημαντική είναι η ύπαρξη ενός λεξικού το οποίο περιέχει όλες τις λέξεις που χρησιμοποιεί το σύστημα και τις μεταγραφές τους. Με αυτόν τον τρόπο το λεξικό ορίζει ποιοι συνδυασμοί φωνημάτων έχουν νόημα βοηθώντας έτσι στην αναζήτηση. Επιπρόσθετα, εισάγεται και μια γραμματική $P(W)$, που ορίζει ποιοι συνδυασμοί λέξεων βγάζουν νόημα και η πληροφορία αυτή ενσωματώνεται στην αναζήτηση Viterbi.

Τέλος, η εκπαίδευση των μοντέλων γίνεται, όπως αναφέρθηκε στην Ενότητα 4.1.1, με επανεκτιμήσεις των παραμέτρων των HMMs με τη χρήση του αλγόριθμου Baum-Welch, ενώ κατά την αποκωδικοποίηση το ακουστικό μοντέλο με το λεξιλόγιο και το γλωσσικό μοντέλο, συνδυάζονται σε ένα ενιαίο δίκτυο πάνω στο οποίο ο αλγόριθμος Viterbi βρίσκει το πιο πιθανό μονοπάτι καταστάσεων (αλληλουχίες λέξεων) με βάση τα δεδομένα παρατηρήσεων τα οποία είναι τα διανύσματα χαρακτηριστικών.

Αναγνώριση Συναισθήματος

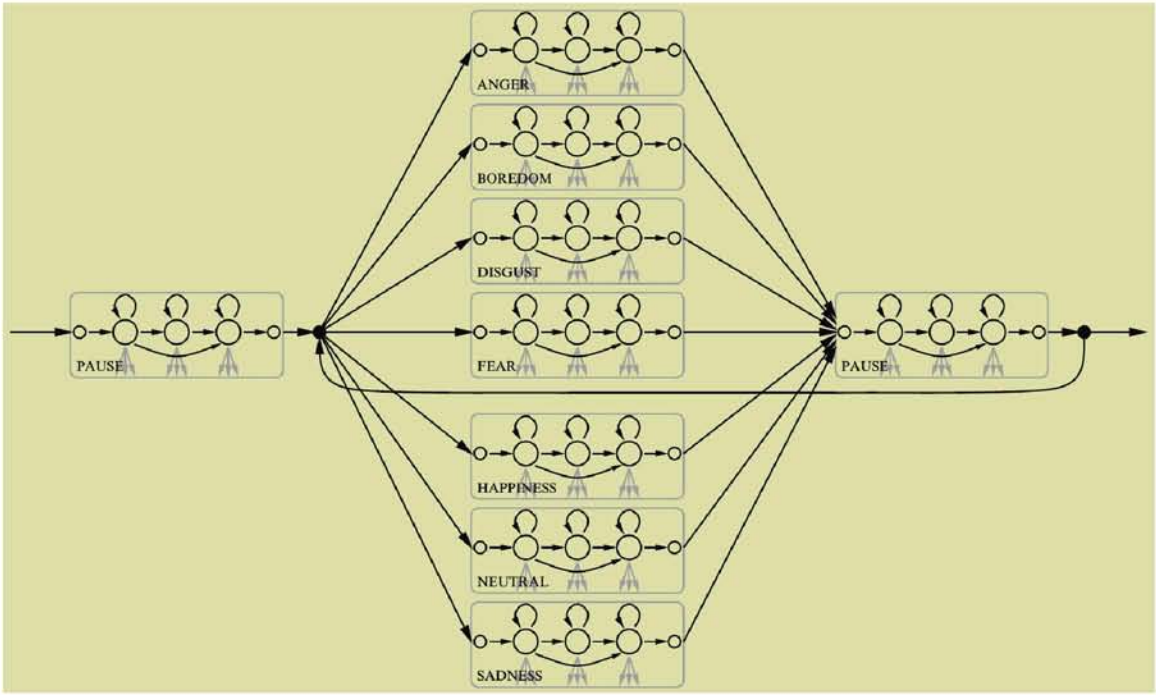
Για τη δημιουργία ενός συστήματος που αναγνωρίζει συναισθήματα χρησιμοποιούνται τα ίδια δομικά συστατικά, δηλαδή το ακουστικό μοντέλο, το λεξιλόγιο και η γραμματική. Τα αρχικά δεδομένα επιβάλλεται να έχουν ετικέτες που δηλώνουν τι συναίσθημα εκφράζει κάθε ηχογράφηση. Μπορεί να υπάρχουν πολλές ετικέτες για κάθε αρχείο σε περίπτωση που παρατηρούνται εναλλαγές συναισθημάτων στην ομιλία, ή μία ετικέτα που χαρακτηρίζει όλη την έκφραση.

Το λεξιλόγιο έχει την ίδια δομή, με τη διαφορά ότι οι λέξεις που περιέχει είναι οι ετικέτες των συναισθημάτων που πρόκειται να αναγνωρίσει το σύστημα. Οι μεταγραφές σε αυτή την περίπτωση είναι ένα φώνημα για κάθε κλάση συναισθήματος.

Στη συνέχεια, απαραίτητη είναι η ύπαρξη μιας γραμματικής συναισθημάτων, η οποία ορίζει ποιες αλληλουχίες συναισθημάτων επιτρέπονται μέσα σε κάθε έκφραση, αλλά και το αν θα υπάρχουν παύσεις ενδιάμεσα. Οι περιορισμοί που βάζει μια τέτοια γραμματική εξαρτώνται από τη δομή των δεδομένων.

Αφού έχουν δημιουργηθεί το λεξιλόγιο και η γραμματική, τα βήματα που ακολουθούν δεν διαφέρουν με αυτά της αναγνώρισης ομιλίας. Η μοντελοποίηση των HMMs γίνεται με τρόπο ώστε ένα φώνημα συναισθήματος να αντιπροσωπεύεται από ένα μοντέλο Markov, το οποίο μπορεί να συνοδεύεται από μοντέλα μίξης όπως παραπάνω.

Τέλος, ο συνδυασμός των μοντέλων μαζί με τη γραμματική και το λεξιλόγιο δημιουργούν ένα δίκτυο παρόμοιο με αυτό της Εικόνας 4.6. Οι μέθοδοι εκπαίδευσης κι επανεκτίμησης των παραμέτρων των μοντέλων συναισθημάτων παραμένουν οι ίδιοι, όπως επίσης και η αναγνώριση γίνεται με τον ίδιο τρόπο πάνω στο δίκτυο συναισθημάτων για την εύρεση του καλύτερου μονοπατιού, δηλαδή για την εύρεση της πιο πιθανής κλάσης συναισθήματος.



Σχήμα 4.6: Δίκτυο αναγνώρισης συναισθημάτων [33]

4.3 Το εργαλείο Kaldi

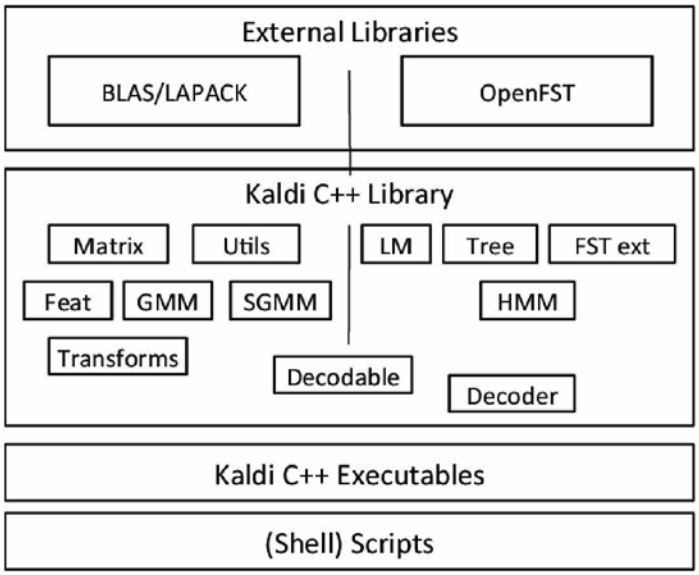
Το βασικό εργαλείο που χρησιμοποιείται σε αυτήν την εργασία τόσο για την δημιουργία και εκπαίδευση των ταξινομητών, όσο και για την εξαγωγή χαρακτηριστικών, είναι το Kaldi. Στην ενότητα αυτή γίνεται παρουσίαση του εργαλείου και η περιγραφή όλης της διαδικασίας, από την προετοιμασία των δεδομένων μέχρι την εκπαίδευση των μοντέλων.

4.3.1 Τι είναι το Kaldi

Το Kaldi [34] είναι ένα εργαλείο ανοιχτού κώδικα γραμμένο σε γλώσσα C++ και χρησιμοποιείται κυρίως σε εφαρμογές αναγνώρισης ομιλίας. Το 2012 κυκλοφόρησε για πρώτη φορά ως ένα εργαλείο ομιλίας γενικού σκοπού. Το κύριο πρόσωπο πίσω από τη δημιουργία του είναι ο Daniel Povey.

Η Εικόνα 4.7 συνοψίζει τα βασικά χαρακτηριστικά που διακρίνουν το Kaldi και επιγραμματικά είναι:

- Χρήση ενός είδους μηχανών πεπερασμένης κατάστασης, τα Finite State Transducers (FSMs).
- Υποστηρίζει μία εκτενή βιβλιοθήκη ρουτινών γραμμικής άλγεβρας των πακέτων LAPACK και BLAS.
- Πληθώρα αλγορίθμων για τη δημιουργία ακουστικών μοντέλων αναγνώρισης, την εφαρμογή μοντέλων Markov, δέντρων απόφασης, νευρωνικών δικτύων καθώς και αλγόριθμους επεξεργασίας δεδομένων.
- Ολοκληρωμένες εφαρμογές μοντέλων αναγνώρισης με ποικιλία υλοποιήσεων και χρήση διαφορετικών βάσεων δεδομένων.

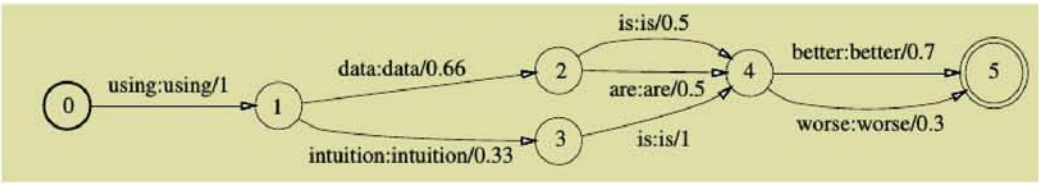


Σχήμα 4.7: Δομικά στοιχεία του Kaldi [34]

4.3.2 Finite State Transducers

Το Kaldi είναι ένα εργαλείο αναγνώρισης ομιλίας το οποίο βασίζεται στα Weighted Finite Transducers (WFSTs), δηλαδή στα FSTs τα οποία έχουν βάρη στις μεταβάσεις μεταξύ των καταστάσεών τους. Η χρήση τους στο πεδίο αναγνώρισης ομιλίας οφείλεται στο οτι παρέχουν μια φυσική και κοινή αναπαράσταση των στοιχείων του συστήματος

αναγνώρισης ομιλίας, δηλαδή των HMMs, του Λεξικού και της Γραμματικής. Όπως φαίνεται στην Εικόνα 4.8, τα WFSTs αποτελούνται από ένα σύνολο καταστάσεων, μία αρχική και μία τελική, και μεταβάσεις μεταξύ τους. Η κάθε μετάβαση απεικονίζεται με ένα σύμβολο εισόδου, ένα σύμβολο εξόδου και ένα βάρος-πιθανότητα. Το όνομα transducer, που σημαίνει μετατροπέας, φανερώνει την λειτουργία των δομών αυτών, η οποία είναι η αντιστοίχηση των συμβόλων εισόδου με αυτά της εξόδου. Με αυτόν τον τρόπο, ένα WFST που περιγράφει το λεξικό ενός συστήματος, για παράδειγμα, έχει ως είσοδο φωνήματα και ως έξοδο την λέξη την οποία σχηματίζουν.



Σχήμα 4.8: Παράδειγμα ενός WFST [35]

Όπως θα δούμε στην επόμενη ενότητα, στόχος είναι η δημιουργία ενός γραφήματος αποκωδικοποίησης το οποίο περιγράφεται από ένα συνδυασμό WFSTs και η εύρεση ενός μονοπατιού Viterbi. Τα βασικά WFSTs που δημιουργούνται από το εργαλείο Kaldi είναι το μοντέλο της Γραμματικής G, το μοντέλο του Λεξικού L, το μοντέλο C συμφραζομένων και το μοντέλο για το HMM H. Το γράφημα αποκωδικοποίησης είναι στην ουσία η σύνθεση $H \circ C \circ L \circ G$ όπου η διαδικασία κατασκευής του περιγράφεται αναλυτικά στο [35].

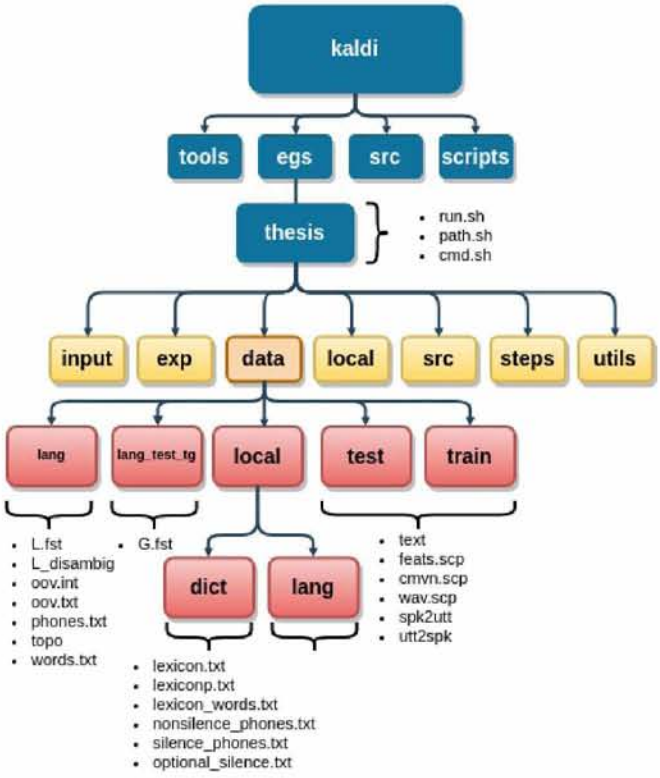
4.3.3 Η γενική εικόνα

Σε αυτή την ενότητα περιγράφονται τα βήματα για τη δημιουργία των ταξινομητών. Τα βήματα περιλαμβάνουν:

- Προετοιμασία Δεδομένων.
- Δημιουργία γραμματικής και του λεξικού.

- Εξαγωγή χαρακτηριστικών MFCC.
- Εκπαίδευση και αποκωδικοποίηση μοντέλου GMM-HMM.
- Εκπαίδευση και αποκωδικοποίηση μοντέλου DNN-HMM.

Μια διευκρίνηση που πρέπει να γίνει, είναι ότι στην περίπτωση των συναισθημάτων θα δημιουργήσουμε ένα διαφορετικό λεξικό και γραμματική από αυτά που χρησιμοποιούνται στην αυτόματη αναγνώριση ομιλίας [36], όπως περιγράφει η ενότητα 4.2. Έτσι, στην παρούσα εργασία προσεγγίζουμε τα συναισθήματα ως λέξεις ή φωνήματα και δημιουργούμε ένα μοντέλο γλώσσας (γραμματική) το οποίο επιτρέπει την αναγνώριση ενός μόνο συναισθήματος από κάθε πρόταση. Η Εικόνα 4.9 δείχνει συνολικά τη δομή και τα αρχεία που πρέπει να δημιουργηθούν.



Σχήμα 4.9: Δομή φακέλων Kaldi

Το αρχείο **run.sh**, η βασική δομή του οποίου μπορεί να μελετηθεί εδώ, είναι η ραχοκοκαλιά της εργασίας, γιατί περιλαμβάνει τις κλήσεις όλων των scripts για το στήσιμο

και την εκπαίδευση του συστήματος, ενώ οι φάκελοι **utils** και **steps** περιέχουν αυτοματοποιημένα scripts με εργαλεία όπως, εκπαίδευση HMM μοντέλων, εξαγωγή χαρακτηριστικών και αποκωδικοποίηση. Οι φάκελοι **input**, **local**, **test**, **train**, καθώς και ο φάκελος των δεδομένων, είναι αυτοί που χρειάζεται να επεξεργαστεί και να δημιουργήσει ο χρήστης. Στο φάκελο **local** βρίσκονται όλα τα scripts (bash, python) που χρειάστηκε να γράψουμε και έχουν να κάνουν κυρίως με προετοιμασία δεδομένων. Το **input** περιλαμβάνει αρχεία που περιγράφουν το γλωσσικό μοντέλο. Τέλος, οι φάκελοι exp και mfcc δημιουργούνται αυτόματα κατά τη διάρκεια της εκτέλεσης. Ο πρώτος περιλαμβάνει τα αρχεία καταγραφής, ενώ ο δεύτερος περιλαμβάνει μητρώα πινάκων με τα εξαγόμενα χαρακτηριστικά των αρχείων ήχου.

4.3.4 Προετοιμασία Δεδομένων

Υπεύθυνο script για την προετοιμασία των δεδομένων είναι το **prepare_data.sh**. Η πρώτη εργασία είναι ο διαχωρισμός των δεδομένων σε ένα σύνολο εκπαίδευσης και ένα σύνολο αξιολόγησης με αναλογία 80%/20% αντίστοιχα. Στη συνέχεια, δημιουργούνται τέσσερα αρχεία για κάθε ένα από τα σύνολα εκπαίδευσης και αξιολόγησης και αποθηκεύονται στους φακέλους **train** και **test**. Είναι αρχεία κειμένου τα οποία επιτρέπουν στο Kaldi να επικοινωνεί με τα αρχεία ήχου. Δημιουργήθηκαν μέσω δικών μας scripts γραμμένα σε γλώσσα python. Αυτά είναι:

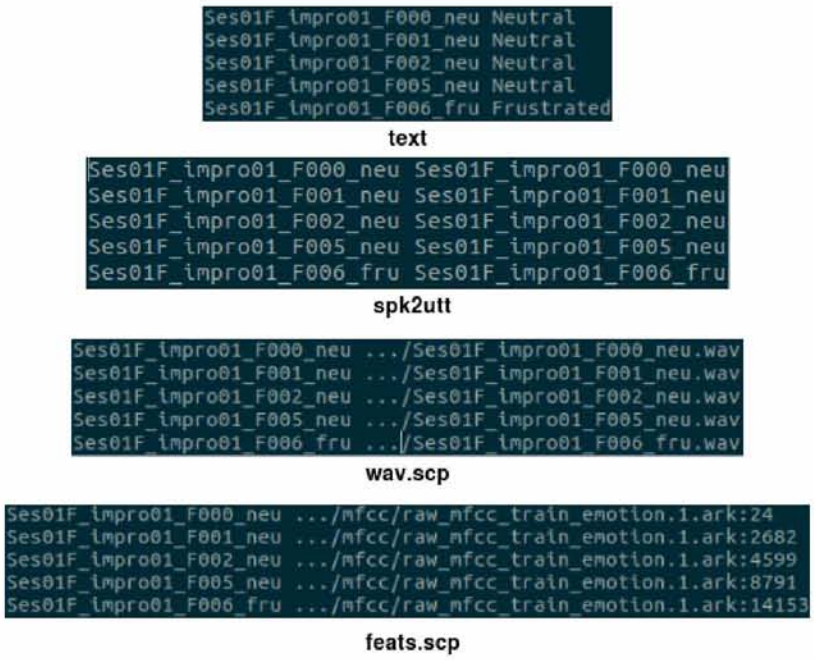
- **text** <utteranceID> <text_transcription>: Περιέχει αντιστοιχίσεις του ονόματος του κάθε αρχείου με τη μεταγραφή του. Για το πρόβλημα της αναγνώρισης συναισθήματος, οι μεταγραφές δηλώνουν την κλάση του συναισθήματος που ανήκει η κάθε ηχογράφηση.

- **wav.scp** <utteranceID> <full_path_to_audio_file> Περιέχει πληροφορία για την τοποθεσία των wav αρχείων. Συνδέει ένα string ID με κάθε αρχείο ήχου.

- **spk2utt** <speakerID> <utteranceID> Αντιστοιχίζει το id του ομιλητή με το id μιας έκφρασης. Στα πλαίσια της εργασίας αγνοήσαμε την ταυτότητα του ομιλητή. Σύμφωνα με την επίσημη ιστοσελίδα του Kaldi, στην περίπτωση αυτή η ταυτότητα του ομιλητή είναι το όνομα του αρχείου.

- **utt2spk** <utteranceID> <speakerID> Το αντίστροφο του spk2utt.

Μέρος των αρχείων αυτών φαίνονται στην Εικόνα 4.10, για μια πιο πλήρη εικόνα της διαδικασίας.



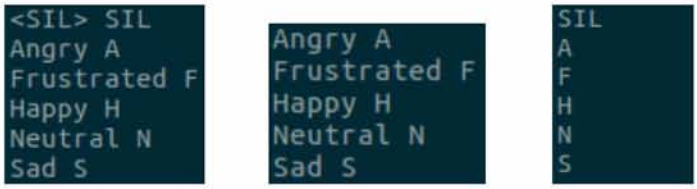
Σχήμα 4.10: Περιεχόμενα βασικών αρχείων κατά την προετοιμασία δεδομένων.

4.3.5 Λεξικό και Γραμματική

Επόμενη ενέργεια είναι η δημιουργία του Λεξικού. Το Λεξικό κανονικά συνδέει τις λέξεις με τις προφορές τους, αλλά στην περίπτωσή μας έχουμε ορίσει ως λέξεις τις κλάσεις των συναισθημάτων, οι οποίες ερμηνεύονται από ένα φώνημα. Το script **prepare_dict.sh** και **prepare_lang** είναι υπεύθυνα για τη δημιουργία του λεξικού που περιγράφεται με ένα FST (L.fst). Για να μπορέσουν να εκτελεστούν τα παραπάνω scripts χρειάζεται να

φτιάξει ο χρήστης τρία αρχεία με προορισμό αποθήκευσης τον φάκελο **input**. Αυτά είναι:

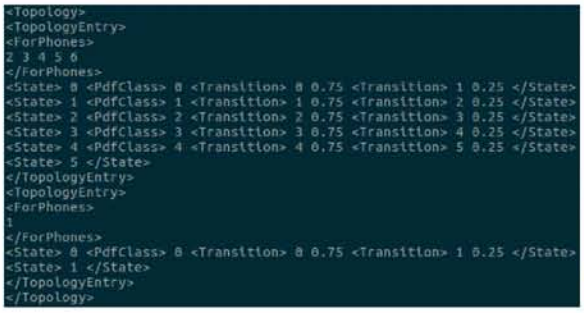
- **lexicon.txt**: Λίστα με όλες τις λέξεις του λεξιλογίου και τις αντίστοιχες προφορές τους. Εδώ κάθε λέξη/κλάση συναισθήματος συνδέεται με ένα φώνημα.
- **lexicon_nosil.txt**: Λίστα με τα μη σιωπηλά φωνήματα.
- **phones.txt**: Λίστα όλων των φωνημάτων του λεξιλογίου.



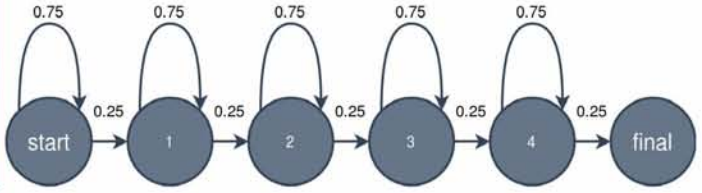
Σχήμα 4.11: Αρχεία για τη δημιουργία μοντέλου γλώσσας.

Όπως φαίνεται στην εικόνα 4.11, οι λέξεις μας είναι τα ονόματα των 5 κατηγοριών συναισθημάτων. **Angry, Frustrated, Happy, Neutral, Sad** και τα αντίστοιχα φωνήματα **A, F, H, N, S**. Με βάση αυτά δημιουργείται ο φάκελος **dict**, μέσω του **prepare_dict.sh**, και περιέχει την ίδια πληροφορία με τα αρχεία που αναφέραμε, όπως φαίνεται στο Σχήμα 4.9. Το **prepare_lang.sh** δέχεται ως είσοδο τα περιεχόμενα του **dict** και δημιουργεί τον φάκελο **data/lang**, ο οποίος περιέχει το FST του λεξικού. Επιπρόσθετα, δημιουργεί το αρχείο **data/lang/topo**, ένα αρχείο κειμένου που περιγράφει την τοπολογία των HMMs που θα χρησιμοποιηθούν. Ο αριθμός των καταστάσεων δίνεται σαν όρισμα κατά την κλήση του **prepare_lang.sh** από τον ίδιο το χρήστη. Η Εικόνα 4.12α΄ παρουσιάζει τα περιεχόμενα του αρχείου **topo**, ενώ στην Εικόνα 4.12β΄ φαίνεται το αντίστοιχο γράφημα HMM.

Αναλυτικότερα, για κάθε φώνημα και κατ΄ επέκταση για κάθε κλάση συναισθήματος, δημιουργείται ένα HMM με τη συγκεκριμένη δομή. Το ίδιο συμβαίνει και για τα φωνήματα ησυχίας. Η πρώτη κατάσταση ορίζεται ως αρχική, ενώ η τελευταία είναι η τελική και



(α′) Τοπολογία HMM για κάθε φώνημα.



(β′) Γράφημα HMM που περιγράφει το αρχείο topo.

Σχήμα 4.12

δεν έχει καμία έξοδο. Η πρώτη μαζί με τις εσωτερικές καταστάσεις χαρακτηρίζονται από μεταβάσεις, είτε σε επόμενη κατάσταση είτε στην ίδια, καθώς και από τις εξόδους που εμφανίζουν τις παρατηρήσεις/χαρακτηριστικά. Τέλος, οι τιμές που δίδονται στις πιθανότητες μετάβασης δεν είναι μόνιμες, αλλά χρησιμοποιούνται μόνο για την αρχικοποίηση του μοντέλου πριν ξεκινήσει η διαδικασία της εκπαίδευσης.

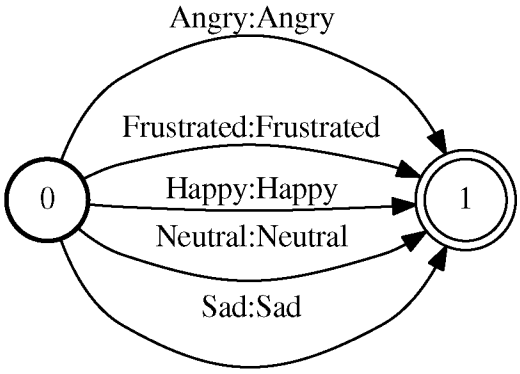
Για τη δημιουργία της Γραμματικής ορίζουμε το δικό μας μοντέλο (G.txt) σε μορφή κειμένου (Πίνακας 4.1) κι έπειτα το μετατρέπουμε σε FST (Εικόνα 4.13) με τη βοήθεια της εντολής της βιβλιοθήκης OpenFst:

```
fstcompile -isymbols=words.txt -osymbols=words.txt G.fst.txt G.fst
```

Από	Προς	Είσοδος	Έξοδος	Βάρος
0	1	Angry	Angry	0.0
0	1	Frustrated	Frustrated	0.0
0	1	Happy	Happy	0.0
0	1	Neutral	Neutral	0.0
0	1	Sad	Sad	0.0
1	0.0			

Πίνακας 4.1: Τοπολογία Γραμματικής σε μορφή κειμένου.

Όπως τονίσαμε, η γραμματική μας επιτρέπει μία λέξη/κλάση για μία πρόταση.



Σχήμα 4.13: Το FST της Γραμματικής.

4.3.6 Εξαγωγή Χαρακτηριστικών

Στο Κεφάλαιο 3 έγινε περιγραφή για τη φύση των χαρακτηριστικών MFCC και τον τρόπο με τον οποίο εξάγονται. Με παρόμοιο τρόπο εξάγονται και στο Kaldi. Το script **run.sh**, μετά τη δημιουργία του Λεξικού και της Γραμματικής, εκτελεί το step/make_mfcc.sh με το οποίο εξάγει τα χαρακτηριστικά MFCC. Αυτά αποθηκεύονται στον φάκελο **mfcc** ως αρχεία .ark, όπου σύμφωνα με την ορολογία του Kaldi, είναι πίνακες χαρακτηριστικών μεγέθους (**N-πλαίσια x M-MFCCs**) για κάθε αρχείο ήχου. Η μορφή τους είναι:

```
Ses01F_impro01_F000_neu [
69.9251 -5.360498 -10.21452 6.206154 -21.335 3.544365 -14.39622 4.282937 -
19.5864 -9.607042 2.231804 6.724803 -1.034574
69.6925 -3.366363 -16.90947 9.380353 -16.899 -3.01285 -22.2214 -1.468073 -
23.0513 -4.671732 -2.535166 7.54028 -10.40728
. . .
]
```

Επιπλέον, μέσα από αυτή τη διαδικασία δημιουργείται και το αρχείο **feats.scp**, το οποίο αποθηκεύεται στους φακέλους **test** και **train**. Η δομή του είναι:

- **feats.scp** <utteranceID> <full_path_to_feature_file> Περιέχει πληροφορία για την τοποθεσία των εξαγόμενων χαρακτηριστικών. Αντιστοιχίζει κάθε πρόταση με το αρχείο χαρακτηριστικών της.

Παρόμοια δομή παρουσιάζει και το αρχείο **cmvn.scp**. Μετά την εξαγωγή χαρακτηριστικών του κάθε ομιλητή, γίνονται αντίστοιχες κανονικοποιήσεις πάνω στα χαρακτηριστικά αυτά. Περιλαμβάνουν κανονικοποιήσεις του μέσου και της διασποράς και συμβάλλουν στη μείωση του θορύβου των ηχογραφήσεων [37].

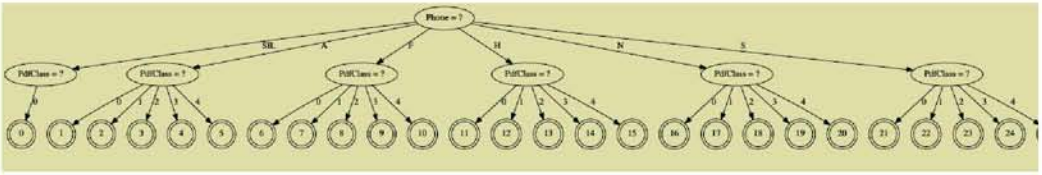
Ο έλεγχος των παραμέτρων της διαδικασίας γίνεται μέσω ενός αρχείου ρυθμίσεων, με το χρήστη να μπορεί να επιλέξει τις τιμές για το μέγεθος του πλαισίου, το μέγεθος της μετατόπισης και το είδος της συνάρτησης παραθύρου. Διαφορετικά, χρησιμοποιούνται οι προκαθορισμένες τιμές. Σημαντική παρατήρηση είναι πως οι συντελεστές delta, delta-delta δεν υπολογίζονται σε αυτό το σημείο. Ο υπολογισμός τους γίνεται στο στάδιο της εκπαίδευσης.

4.3.7 Εκπαίδευση GMM-HMM

Η βασική εντολή για την εκπαίδευση του μοντέλου GMM-HMM, είναι η **train_mono.sh** και περιλαμβάνει τα εξής στάδια:

1. **Αρχικοποίηση**: Δημιουργείται το αρχικό μοντέλο GMM-HMM μαζί με το αντίστοιχο δέντρο απόφασης (Εικόνα 4.14). Επιπλέον, για κάθε έκφραση του συνόλου εκπαίδευσης, δημιουργούνται FSTs που αντιστοιχούν σε γραφήματα *HCLG*, τα οποία θα χρησιμοποιηθούν κατά την εκπαίδευση και για αυτό το λόγο οι πιθανότητες μετάβασης δεν έχουν οριστεί ακόμη. Επίσης, σε αυτό το στάδιο της αρχικοποίησης κάθε κατάσταση του μοντέλου περιγράφεται από ένα μόνο Γκαουσιανό στοιχείο.

2. **Στάδιο εκπαίδευσης:** Για την εκκίνηση της εκπαίδευσης, όπως αναφέρθηκε στην Ενότητα 4.1.1, ορίζεται μια αρχική ευθυγράμμιση των καταστάσεων με



Σχήμα 4.14: Δέντρο απόφασης.

τα διανύσματα χαρακτηριστικών, όπου κάθε κατάσταση καλύπτει τον ίδιο αριθμό παρατηρήσεων. Στο σημείο αυτό πρέπει να αναφερθεί ότι ο αλγόριθμος που χρησιμοποιείται είναι ο Viterbi. Μετά την πρώτη επανάληψη (αρχική ευθυγράμμιση), συνεχίζονται οι επαναληπτικές διορθώσεις των παραμέτρων, όπως επιβάλλει ο αλγόριθμος, κι επίσης αυξάνεται ο αριθμός των Γκαουσιανών στοιχείων σε κάθε κατάσταση.

Εδώ πρέπει να σημειωθεί πως οι συντελεστές delta και delta-delta, προστίθενται στα διανύσματα των χαρακτηριστικών σε αυτό το τμήμα του κώδικα, δηλαδή ακριβώς πριν αρχίσει η διαδικασία της εκπαίδευσης.

4.3.8 Αποκωδικοποίηση GMM-HMM

Πριν από τη διαδικασία της αποκωδικοποίησης, δημιουργείται ένα γράφημα αποκωδικοποίησης WFST με την εκτέλεση του **mkgraph.sh** που περιγράφεται από τη σύνθεση $H \circ C \circ L \circ G$. Η διαδικασία της αποκωδικοποίησης ξεκινάει με την εκτέλεση της εντολής **decode.sh**. Η ιδέα είναι ότι το γράφημα $HCLG$ πλέον περιγράφει το μοντέλο ως μια σειρά από GMMs και από μεταβάσεις μεταξύ των καταστάσεων. Έτσι, η αποκωδικοποίηση περιλαμβάνει την εύρεση του μονοπατιού Viterbi στο γράφημα αυτό.

4.3.9 Εκπαίδευση DNN-HMM

Για την δημιουργία και εκπαίδευση ενός μοντέλου DNN-HMM είναι απαραίτητη η ύπαρξη των ευθυγραμμίσεων που δημιουργήθηκαν στα προηγούμενα βήματα, γιατί από αυτές θα ξεκινήσει η εκπαίδευση των νευρωνικών δικτύων. Εδώ φαίνεται πόσο σημαντικό είναι να

αποδίδει το μοντέλο GMM-HMM που έχει δημιουργηθεί. Το Kaldi διαθέτει τρία είδη υλοποιήσεων βαθιών νευρωνικών δικτύων. Η πρώτη ονομάζεται nnet1 και χρησιμοποιεί "Sequence-discriminative training" [38] για τα δίκτυά της. Πλέον δεν υποστηρίζεται λόγω παλαιότητας, με τις nnet2 και nnet3 να είναι οι βασικότερες υλοποιήσεις που χρησιμοποιούνται από την κοινότητα του Kaldi. Το νευρωνικό δίκτυο αυτής της εργασίας είναι της υλοποίησης nnet2 [39], με τα βασικά χαρακτηριστικά του να παρουσιάζονται αναλυτικά στο [40]. Συνοπτικά, για συνάρτηση κόστους χρησιμοποιείται η συνάρτηση cross-entropy και ο αλγόριθμος για τη βελτιστοποίησή της είναι ο Stohastic Gradient Descent (SGD) με οπισθοδιάδοση. Επιπλέον, για τα ενδιάμεσα επίπεδα χρησιμοποιείται η συνάρτηση ενεργοποίησης p-norm:

$$y = ||x||_p = (\sum_{i=1} |x_i|^p)^{1/p}, \tag{4.14}$$

ενώ για το επίπεδο εξόδου η συνάρτηση soft-maxout με τύπο:

$$y = \log \sum_{i=1} e^{x_i}. \tag{4.15}$$

Για την έναρξη της εκπαίδευσης εκτελείται το script, **train_pnorm_fast.sh**. Ο χρήστης έχει τη δυνατότητα να περάσει μια πληθώρα τιμών ως ορίσματα για την παραμετροποίηση της διαδικασίας. Μερικά από τα πιο βασικά περιλαμβάνουν, τον αριθμό των epochs, τον αριθμό των κρυφών επιπέδων του δικτύου και τον ρυθμό μάθησης. Κατά την αρχικοποίηση του σταδίου αυτού ξεκινάμε με ένα κρυφό επίπεδο και σιγά σιγά προστίθενται καινούργια όσο προχωράει η διαδικασία, μέχρι τον επιθυμητό αριθμό επιπέδων που έχει ορίσει ο χρήστης. Επιπρόσθετα, όσο διαρκεί η εκπαίδευση ο ρυθμός μάθησης μειώνεται σταδιακά μέχρι που σταθεροποιείται για τα τελευταία epochs. Για την ολοκλήρωση της εκπαίδευσης, στην τελευταία λαμβάνεται υπόψιν όχι μόνο το αποτέλεσμα της τελευταίας επανάληψης, αλλά και των n (ορισμένο από τον χρήστη) προηγούμενων, κάτι το οποίο δίνει σημαντική βελτίωση στην απόδοση [40].

Κεφάλαιο 5

Πειράματα και Αποτελέσματα

5.1 Επισκόπηση

Τα πειράματα που ακολουθούν εκτελέστηκαν σε τοπικό επίπεδο, σε προσωπικό υπολογιστή με τα εξής χαρακτηριστικά:

- NVIDIA GeForce GTX 1060 6GB
- Intel Core i5 7600K
- 8GB RAM
- 2TB HDD Storage
- 120GB SSD Storage

Όσον αφορά στα δεδομένα, χρησιμοποιήθηκαν συνολικά 7380 ηχογραφημένες εκφράσεις της συλλογής δεδομένων IEMOCAP οι οποίες φέρουν μία ετικέτα συναισθήματος η καθεμιά. Συνολικά υπάρχουν πέντε κλάσεις συναισθημάτων, χαρά, λύπη, θυμός, ουδέτερο, εκνευρισμός. Οι εκφράσεις χωρίστηκαν σε δύο σύνολα, το σύνολο εκπαίδευσης που αποτελεί το 80% του συνολικού αριθμού των εκφράσεων και το σύνολο αξιολόγησης που περιέχει το υπόλοιπο 20%. Το βασικό εργαλείο που χρησιμοποιήθηκε

τόσο για την εξαγωγή χαρακτηριστικών όσο και για την εκπαίδευση των μοντέλων είναι το Kaldi. Από κάθε ηχογράφηση-έκφραση προέκυψαν 39 MFCCs (13 MFCCs + 13 Delta + 13 Delta-Delta) για κάθε πλαίσιο.

5.2 Μετρικές αξιολόγησης

Προτού παρουσιαστούν τα αποτελέσματα των μοντέλων αναγνώρισης, είναι αναγκαίο να οριστεί ο τρόπος με τον οποίο αξιολογείται η απόδοσή τους. Αυτό γίνεται με το ποσοστό σφάλματος λέξης ή Word Error Rate (WER), μια από τις πιο γνωστές μετρικές στον κλάδο της Αυτόματης Αναγνώρισης Ομιλίας [41], και εκφράζει το ποσοστό λάθους κατά την αναγνώριση λέξεων που στην περίπτωσή μας ισοδυναμεί με ανγνώριση συναισθήματος πρότασης. Το ποσοστό αυτό υπολογίζεται από:

$$WER = \frac{S + D + I}{N} \cdot 100\%, \tag{5.1}$$

όπου S είναι ο αριθμός των αντικαταστάσεων, I είναι ο αριθμός των εισαγωγών, D είναι ο αριθμός των διαγραφών και N ο συνολικός αριθμός των παρατηρήσεων. Από αυτούς τους αριθμούς προκύπτουν αντίστοιχα οι παρακάτω μετρικές:

- Λάθος εισαγωγής. Το λάθος που υπολογίζεται όταν ο ταξινομητής αναγνωρίζει περισσότερες λέξεις από όσες έπρεπε.
- Λάθος διαγραφής. Όταν ο ταξινομητής αναγνωρίζει μία λέξη ενώ δεν έχει ειπωθεί.
- Λάθος αντικατάστασης. Όταν αναγνωρίζεται η λάθος λέξη στη θέση κάποιας άλλης.

5.3 Αποτελέσματα

Σε αυτή την υποενότητα θα παρουσιαστούν οι αποδόσεις των δύο ταξινομητών αυτής της εργασίας. Και τα δύο συστήματα χαρακτηρίζονται από πολλές παραμέτρους, επομένως είναι σημαντικό να προσδιοριστεί για ποιές τιμές καταλήγουμε στο καλύτερο δυνατό αποτέλεσμα και γιατί.

5.3.1 Αποτελέσματα GMM-HMM

Όπως έχει αναφερθεί, το βασικό στοιχείο για τη δημιουργία του τελικού ταξινομητή είναι η ύπαρξη ενός μοντέλου GMM-HMM από το οποίο αντλεί πληροφορία. Κύριο ρόλο διαδραματίζουν οι παράμετροι που ελέγχουν τον αριθμό των καταστάσεων για τα μοντέλα Markov και τον αριθμό των Γκαουσιανών στοιχείων του μοντέλου. Ο σωστός προσδιορισμός των τιμών τους επηρεάζει όχι μόνο την αποτελεσματικότητα του ταξινομητή αλλά και τον χρόνο εκπαίδευσης. Είναι εύλογο ένα μοντέλο με πολλές καταστάσεις να χρειάζεται πολύ περισσότερο χρόνο εκπαίδευσης και αποκωδικοποίησης. Στα πειράματά μας, τα οποία φαίνονται αναλυτικά στον Πίνακα 5.1, παρατηρήθηκε πως το σύστημα έδωσε τα καλύτερα αποτελέσματα για μικρό αριθμό καταστάσεων και για την περίπτωση των σιωπηλών, αλλά και για των μη σιωπηλών. Αυτό συμβαίνει διότι έχουμε να κάνουμε με μικρό αριθμό φωνημάτων, ειδικά για την περίπτωση του μοντέλου ησυχίας. Το τελικό μοντέλο αποτελείται από 5 καταστάσεις για τα φωνήματα μη ησυχίας ή αλλιώς, 5 καταστάσεις για μία κλάση συναισθήματος (βλ. Ενότητα 4.2), και μία κατάσταση για το φώνημα της ησυχίας. Επιπλέον, παρατηρήθηκε ότι μια καλή τιμή για τον μέγιστο συνολικό αριθμό των Γκαουσιανών στοιχείων του μοντέλου είναι 1000.

Παρατηρούμε πως το ποσοστό λάθους στην καλύτερη περίπτωση βρίσκεται κοντά στο 50%, δηλαδή από το σύνολο των προτάσεων ο ταξινομητής εντοπίζει σωστά το συναίσθημα που εκφράζουν μόνο στις μισές από αυτές. Με βάση αυτό το μοντέλο δημιουργήθηκε και η υλοποίηση με το βαθύ νευρωνικό δίκτυο.

Μη σιωπηλές καταστάσεις	Σιωπηλές καταστάσεις	Γκαουσιανά στοιχεία	WER(%)
5	1	1000	53.45
3	1	1000	55.67
5	3	1000	58.44
4	4	1000	60.98
10	1	2000	53.76
5	1	2000	54.06
5	1	400	58.63
5	3	400	58.72

Πίνακας 5.1: Αποτελέσματα ταξινομητή GMM-HMM

5.3.2 Αποτελέσματα DNN-GMM

Με τη δημιουργία του ταξινομητή DNN-HMM ευελπιστούμε να δούμε βελτίωση των προηγούμενων αποτελεσμάτων και σε τι βαθμό συμβαίνει. Η παραμετροποίηση των νευρωνικών δικτύων δεν είναι εύκολη διαδικασία, διότι τα νευρωνικά δίκτυα είναι πολύπλοκες δομές και εξαρτώνται από πάρα πολλές παραμέτρους.

Όπως είδαμε στο προηγούμενο Κεφάλαιο, ο ταξινομητής που εκπαιδεύτηκε ανήκει στην κατηγορία υλοποιήσεων nnet2 του Kaldi. Οι βασικές παράμετροι που επηρέασαν σημαντικά τα αποτελέσματα είναι ο αριθμός των epochs και το πόσα κρυφά επίπεδα θα χρησιμοποιηθούν. Η υλοποίηση με την καλύτερη απόδοση όπως φαίνεται στον Πίνακα 5.2 διαθέτει δύο κρυφά επίπεδα και 15 epochs και άλλα επιπλέον 5. Είναι εμφανές πως με τον αριθμό δεδομένων που διαθέτουμε η εκπαίδευση με 2 κρυφά επίπεδα είναι αρκετή, καθώς από τα 3 επίπεδα και πάνω το ποσοστό λάθους όλο και αυξάνεται. Επιπρόσθετα, ο ρυθμός εκπαίδευσης που επιλέχθηκε ήταν ο προτεινόμενος, δηλαδή 0.04 για το αρχικό στάδιο εκπαίδευσης και 0.004 για το τελικό.

Εξετάζοντας τα αποτελέσματα διαπιστώνεται μια βελτίωση της απόδοσης του ταξινομητή DNN-HMM σε σχέση με τον GMM-HMM, αλλά είναι πολύ μικρή και αυτό οφείλεται στην αρχική κακή απόδοση του πρώτου μοντέλου, σύμφωνα με την κοινότητα του Kaldi.

Αριθμός Epochs	Κρυφά Επίπεδα	Επιπλέον Epochs	WER(%)
15	2	5	50.30
10	2	5	51.71
8	2	5	51.53
3	2	5	53.19
10	2	2	52.45
15	3	5	51.10
7	3	5	52.20
7	1	5	52.70

Πίνακας 5.2: Αποτελέσματα ταξινομητή DNN-HMM

Κεφάλαιο 6

Σύνοψη

Στην εργασία αυτή εξετάστηκε το πρόβλημα της αναγνώρισης συναισθήματος από ομιλία με τη δημιουργία δύο διαφορετικών μοντέλων ταξινομητών, GMM-HMM και DNN-HMM. Δημιουργήθηκαν μέσω του εργαλείου Kaldi, το οποίο χρησιμοποιείται σε προβλήματα αυτόματης αναγνώρισης ομιλίας. Τα τέσσερα βασικά στοιχεία που απαρτίζουν την έρευνα αυτή είναι η εύρεση κατάλληλων δεδομένων, η εξαγωγή χαρακτηριστικών, και η δημιουργία και εκπαίδευση κατάλληλων μοντέλων ταξινόμησης. Ένα, ίσως προφανές, αλλά ενδιαφέρον πόρισμα που προκύπτει είναι το πόσο σημαντική καθίσταται η συνεργασία των τεσσάρων αυτών βασικών στοιχείων για την επίτευξη του στόχου της έρευνας. Τα στάδια της παρούσας έρευνας συνοψίζονται στα:

- Συλλογή δεδομένων από τη βάση IEMOCAP.
- Εξαγωγή MFCC χαρακτηριστικών.
- Δημιουργία και εκπαίδευση ταξινομητή GMM-HMM.
- Δημιουργία και εκπαίδευση ταξινομητή DNN-HMM.

Επιπλέον, η εργασία αυτή αναδεικνύει τη χρήση του εργαλείου Kaldi σε προβλήματα αναγνώρισης συναισθήματος κάνοντας μια αντιστοίχηση μεταξύ αυτόματης αναγνώρισης ομιλίας και αναγνώρισης συναισθήματος, εφαρμόζοντας ένα λεξικό συναισθημάτων.

Όσον αφορά στην απόδοση των συστημάτων που δημιουργήθηκαν, υπήρξε μια μικρή βελτίωση με τη χρήση νευρωνικών δικτύων αλλά το συνολικό ποσοστό σφάλματος παρέμεινε υψηλό, εξαιτίας του αρχικού ταξινομητή GMM-HMM. Η βελτίωση και εύρεση της αιτίας του σφάλματος μπορεί να αποτελέσει πηγή για νέες έρευνες.

6.1 Μελλοντική Έρευνα

Πολλές είναι οι έρευνες που μπορούν να προκύψουν, και υπάρχουν ήδη πολλές, με αφορμή τη βελτίωση αποτελεσμάτων αναγνώρισης συναισθήματος. Τα θέματά τους μπορούν να αφορούν σε:

- Εξαγωγή πιο εξειδικευμένων χαρακτηριστικών που έχουν τη δυνατότητα να αποσπάσουν τη συναισθηματική κατάσταση του ατόμου που μιλάει. Τέτοια χαρακτηριστικά παρουσιάζονται στην [42]. Επιπλέον, υπάρχουν εργαλεία όπως το [43] που μπορούν και συνδυάζουν χαρακτηριστικά δημιουργώντας νέους συνδυασμούς.

- Μία επίσης διαδεδομένη μέθοδος για την βελτίωση αποτελεσμάτων αναγνώρισης, είναι η χρησιμοποίηση και ο συνδυασμός διαφορετικών πηγών παραγωγής συναισθήματος (φωνή, εκφράσεις προσώπου, κείμενα) [44]. Τα αποτελέσματα τέτοιων ερευνών είναι ικάνα να συμβάλλουν σε μια πιο ακριβέστερη επίλυση τέτοιων προβλημάτων.

Βιβλιογραφία

[1] N. Fragopanagos and J. Taylor, "Emotion recognition in human-computer interaction," *Neural networks : the official journal of the International Neural Network Society*, vol. 18, pp. 389–405, 06 2005.

[2] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, pp. 32 – 80, 02 2001.

[3] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *IJSE*, vol. 1, pp. 68–99, 01 2010.

[4] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, 01 2018.

[5] V. Petrushin, "Emotion in speech: Recognition and application to call centers," *Proceedings of Artificial Neural Networks in Engineering*, 01 2000.

[6] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *INTERSPEECH*, 2014.

[7] M. S. Fahad, J. Yadav, G. Pradhan, and A. Deepak, "Dnn-hmm based speaker adaptive emotion recognition using proposed epoch and mfcc features," *ArXiv*, vol. abs/1806.00984, 2018.

[8] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, "Deep learning based emotion recognition system using speech features and transcriptions," *ArXiv*, vol. abs/1906.05681, 2019.

[9] X. Zhang, Y. Sun, and Shufei Duan, "Progress in speech emotion recognition," in *TENCON 2015 - 2015 IEEE Region 10 Conference*, pp. 1–6, Nov 2015.

[10] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, pp. 1162–1181, 09 2006.

[11] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, pp. 1062–1087, 11 2011.

[12] S. Koolagudi and K. Rao, "Emotion recognition from speech: A review," *International Journal of Speech Technology*, vol. 15, 06 2012.

[13] R. Cowie, E. Douglas-Cowie, S. Savvidou, and E. McMahon, "Feeltrace: an instrument for recording perceived emotion in real time," pp. 19–24, 9 2000. Speech and Emotion: Proceedings of the ISCA workshop; Conference date: 01-09-2000 Through 01-09-2000.

[14] C.-H. Wu, J.-C. Lin, and W.-L. Wei, "Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies," *APSIPA Transactions on Signal and Information Processing*, vol. 3, 11 2014.

[15] D. Ververidis and C. Kotropoulos, "A state of the art review on emotional speech databases," Citeseer, 2003. Proceedings of 1st Richmedia Conference.

[16] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower Provost, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 12 2008.

[17] M. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, pp. 572–587, 03 2011.

[18] G. Andrade-Miranda, *Analyzing of the vocal fold dynamics using laryngeal videos.* PhD thesis, 06 2017.

[19] J. Deller, J. Hansen, and J. Proakis, *Discrete-Time Processing of Speech Signals.* An IEEE Press classic reissue, Wiley, 2000.

[20] K. R. Anne, S. Kuchibhotla, and H. D. Vankayalapati, *Acoustic Modeling for Emotion Recognition.* Springer Publishing Company, Incorporated, 2015.

[21] F. Eyben, *Real-Time Speech and Music Classification by Large Audio Feature Space Extraction.* Springer Publishing Company, Incorporated, 1st ed., 2016.

[22] K. S. Rao and S. G. Koolagudi, "Robust emotion recognition using spectral and prosodic features," in *Springer Briefs in Electrical and Computer Engineering*, 2013.

[23] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. A. Mahjoub, "Speech emotion recognition: Methods and cases study," in *ICAART*, 2018.

[24] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book.* 01 2002.

[25] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, pp. 257–286, Feb 1989.

[26] H. Shimodaira and S. Renals, "Automatic Speech Recognition— ASR Lectures 4&524&28 January 2019," 2019. URL: `http://www.inf.ed.ac.uk/teaching/courses/asr/2018-19/asr03-hmmgmm-handout.pdf`.

[27] F. F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain.," *Psychological review*, vol. 65 6, pp. 386–408, 1958.

[28] D. Hughes and N. Correll, "Distributed machine learning in materials that couple sensing, actuation, computation and communication," *CoRR*, vol. abs/1606.03508, 2016.

[29] M. Nielsen, *Neural Networks and Deep Learning*. Determination Press, 2015.

[30] L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, and H. Sahli, "Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 312–317, Sep. 2013.

[31] J. Niu, Y. Qian, and K. Yu, "Acoustic emotion recognition using deep neural network," in *The 9th International Symposium on Chinese Spoken Language Processing*, pp. 128–132, Sep. 2014.

[32] J. Wagner, T. Vogt, and E. André, "A systematic comparison of different hmm designs for emotion recognition from acted and spontaneous speech," in *Affective Computing and Intelligent Interaction* (A. C. R. Paiva, R. Prada, and R. W. Picard, eds.), (Berlin, Heidelberg), pp. 114–125, Springer Berlin Heidelberg, 2007.

[33] J. Pittermann, A. Pittermann, and W. Minker, *Hybrid Approach to Speech–Emotion Recognition*, pp. 107–149. Dordrecht: Springer Netherlands, 2010.

[34] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society, Dec. 2011. IEEE Catalog No.: CFP11SRW-USB.

[35] M. Mohri, F. Pereira, and M. Riley, *Speech Recognition with Weighted Finite-State Transducers*, pp. 559–584. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[36] A. Pittermann and J. Pittermann, "Getting bored with htk? using hmms for emotion recognition from speech signals," in *2006 8th international Conference on Signal Processing*, vol. 1, Nov 2006.

[37] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1, pp. 133 – 147, 1998.

[38] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2345–2349, 01 2013.

[39] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of deep neural networks with natural gradient and parameter averaging," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015.

[40] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 215–219, May 2014.

[41] R. Errattahi, A. El Hannani, and H. Ouahmane, "Automatic speech recognition errors detection and correction: A review," *Procedia Computer Science*, vol. 128, pp. 32–37, 01 2018.

[42] A. Jacob and P. Mythili, "Prosodic feature based speech emotion recognition at segmental and supra segmental levels," in *2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*, pp. 1–5, Feb 2015.

[43] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, (New York, NY, USA), p. 1459–1462, Association for Computing Machinery, 2010.

[44] G. Castellano, L. Kessous, and G. Caridakis, *Emotion Recognition through Multiple Modalities: Face, Body Gesture, Speech*, pp. 92–103. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.